

LECTURES 1 AND 2

We should start with administrative stuff from the syllabus.

- Instructor: Michael Damron
 - mdamron at indiana dot edu
 - 315 Rawles Hall
 - 812-855-8670
- grader: Yu-yuan Chen (as last semester)
- office hour times
 - Please send me an email with all times during which you cannot make it to office hours.
- Text: same as last semester. Billingsley. Probability and measure, 3rd ed.
- homework due on Wednesday in class. There is no homework due this Wednesday but there will be one due next Wednesday. I will post it on Wednesday.
- Grade breakdown: 40 points homework, 40 points final (take home), 20 points midterm (in class)
- Topics: finish convergence in distribution, ergodic theorem, infinitely divisible laws; conditional expectation, martingales; stochastic processes, Brownian motion. If there is left over time we can discuss additional topics.

1 Weak convergence in $\mathbb{R}^d$

We will begin by discussing extensions of distributional convergence statements from $\mathbb{R}$ to $\mathbb{R}^d$ for $d \geq 2$. This will serve to both cover this material and also review material from last semester. Many statements will not be proved, as they are similar to the ones we showed in $d = 1$.

Definition 1.1. A sequence (μ_n) of (Borel) probability measures on $\mathbb{R}^d$ converges weakly to μ if

$$F_n(x) \rightarrow F(x) \text{ for all continuity points } x \text{ of } F .$$

Here F_n, F are the distribution functions of μ_n, μ :

$$F(x) = F(x_1, \dots, x_d) = \mu \left(\prod_{i=1}^d (-\infty, x_i] \right) .$$

In this case we say $F_n \rightarrow F$ weakly. Further if (X_n) is a sequence of random variables with laws μ_n then X_n converges to X in distribution (and write $X_n \Rightarrow X$) if $\mu_n \rightarrow \mu$ weakly.

Just as in the one-dimensional setting, we have many equivalent conditions for weak convergence.

Theorem 1.2 (Portmanteau in $\mathbb{R}^d$). *Let (μ_n) , μ be probability measures on $\mathbb{R}^d$ with distribution functions (F_n) , F . The following are equivalent.*

1. $\mu_n \rightarrow \mu$ weakly.
2. $\int f \, d\mu_n \rightarrow \int f \, d\mu$ for all bounded continuous $f : \mathbb{R}^d \rightarrow \mathbb{R}$.
3. $\mu_n(A) \rightarrow \mu(A)$ for all Borel $A \subset \mathbb{R}^d$ such that $\mu(\partial A) = 0$.
4. $\limsup_n \mu_n(C) \leq \mu(C)$ for all closed $C \subset \mathbb{R}^d$.
5. $\liminf_n \mu_n(O) \geq \mu(O)$ for all open $O \subset \mathbb{R}^d$.

Proof. The first three statements are analogues of ones we proved were equivalent in one-dimension. You can read the equivalence of these in higher dimensions in Billingsley, Theorem 29.1. The last two are equivalent simply by taking complements. So we will only show that 2 implies 4 and 4 and 5 imply 3.

Assume 4 and 5. For any A with $\mu(\partial A) = 0$, write $\text{int } A$ for the interior of A and $\bar{A}$ for the closure of A . Then

$$\limsup_n \mu_n(\bar{A}) \leq \mu(\bar{A}) = \mu(\text{int } A) + \mu(\partial A) = \mu(\text{int } A) \leq \liminf_n \mu_n(\text{int } A) .$$

However $\mu_n(\bar{A}) \geq \mu_n(\text{int } A)$, for all n , so 3 holds.

To show 2 implies 4 we will approximate the indicator of C by a bounded continuous function. So given $\epsilon > 0$ define the enlarged set

$$C_\epsilon = \{x \in \mathbb{R}^d : \text{dist}(x, C) < \epsilon\} .$$

Now set $f_\epsilon : \mathbb{R}^d \rightarrow [0, 1]$ as the function which equals 1 on C , 0 on C_ϵ^c and is linear in between:

$$f_\epsilon(x) = \begin{cases} 0 & \text{if } x \notin C_\epsilon \\ 1 - \frac{\text{dist}(x, C)}{\epsilon} & \text{if } x \in C_\epsilon \end{cases} .$$

This is a continuous function and because C is closed, you can check that $f_\epsilon(x) \downarrow \mathbf{1}_C(x)$ for all x (for a general C this would converge to the indicator of $\bar{C}$). So under 2, we get

$$\mu_n(C) \leq \int f_\epsilon(x) \, d\mu_n(x) \rightarrow \int f_\epsilon(x) \, d\mu(x) .$$

By monotone convergence, the right converges as $\epsilon \rightarrow 0$ to $\mu(C)$ so we obtain 4. $\square$

From this version we can derive the d -dimensional version of the mapping theorem. Recall that if μ is a probability measure on $\mathbb{R}^d$ and $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ is measurable then the assignment

$$\mu f^{-1}(A) = \mu(f^{-1}(A)) , \quad A \subset \mathbb{R}^k \text{ Borel}$$

defines a probability measure on $\mathbb{R}^k$; this is the push forward of μ through f .

Theorem 1.3 (Mapping theorem). *Let (μ_n) be a sequence of probability measures on $\mathbb{R}^d$ that converge weakly to μ . If $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ write D_f for its set of discontinuities. If $\mu(D_f) = 0$ then $\mu_n f^{-1} \rightarrow \mu f^{-1}$ weakly.*

Proof. We will show that if C is closed then

$$\limsup_n \mu_n(f^{-1}(C)) \leq \mu(f^{-1}(C)) . \quad (1)$$

To do this, first note that

$$\overline{f^{-1}(C)} \subset D_f \cup f^{-1}(C) .$$

Indeed, if $x \in \overline{f^{-1}(C)}$ but $x \notin D_f$, we will show that $x \in f^{-1}(C)$. f is continuous at x , so letting (x_n) be a sequence in $f^{-1}(C)$ which converges to x , we have $f(x_n) \rightarrow f(x)$. Since C is closed and each $f(x_n)$ is in C , we also have $f(x) \in C$; that is, $x \in f^{-1}(C)$.

Since $\mu_n \rightarrow \mu$ weakly,

$$\begin{aligned} \limsup_n \mu_n(f^{-1}(C)) &\leq \limsup_n \mu_n(\overline{f^{-1}(C)}) \\ &\leq \mu(\overline{f^{-1}(C)}) \leq \mu(D_f) + \mu(f^{-1}(C)) = \mu(f^{-1}(C)) , \end{aligned}$$

showing (1). □

As before, this implies that if f is continuous then $\mu_n f^{-1} \rightarrow \mu f^{-1}$ weakly. In other words, if $X_n \rightarrow X$ in distribution and f is continuous then $f(X_n) \rightarrow f(X)$ in distribution.

To finish the general discussion of weak convergence, we recall the notion of tightness.

Definition 1.4. *A sequence (μ_n) of probability measures on $\mathbb{R}^d$ is tight if for each $\epsilon > 0$ there exists a compact set K such that*

$$\mu_n(K) \geq 1 - \epsilon \text{ for all } n .$$

Just as in one dimension,

Theorem 1.5. *If (μ_n) is tight, there exists a probability measure on $\mathbb{R}^d$ and a subsequence (μ_{n_k}) such that $\mu_{n_k} \rightarrow \mu$ weakly.*

2 Multidimensional characteristic functions

In the one dimensional case, we defined the characteristic function of probability measure μ as $\phi_\mu(t) = \int e^{itx} d\mu(x)$. In higher dimensions, we simply replace the product with dot product.

Definition 2.1. *Let μ be a probability measure on $\mathbb{R}^d$. The characteristic function of μ is $\phi_\mu : \mathbb{R}^d \rightarrow \mathbb{R}$ given by*

$$\phi_\mu(t) = \int e^{it \cdot x} d\mu(x) .$$

Here are some properties.

1. The characteristic function uniquely identifies the measure by an inversion formula analogous to the one in one dimension. Precisely, if $A \subset \mathbb{R}^d$ is a rectangle of the form

$$A = \prod_{i=1}^d (a_i, b_i] \text{ with } \mu(\partial A) = 0 ,$$

then

$$\mu(A) = \lim_{T \rightarrow \infty} \frac{1}{(2\pi)^d} \int_{\{|t_j| \leq T \ \forall j\}} \prod_{j=1}^d \frac{e^{-it_j a_j} - e^{-it_j b_j}}{it_j} \phi_\mu(t) \, dt .$$

2. Let $X = (X_1, \dots, X_d)$ be a random vector with characteristic function ϕ_X . There is an idea, called the Cramér-Wold device, to consider $t \cdot X$ as a random variable (one dimensional random vector) for fixed $t \in \mathbb{R}^d$. Its characteristic function is

$$\phi_{t \cdot X}(s) = \mathbb{E} e^{is(t \cdot X)} \text{ for } s \in \mathbb{R} .$$

We claim that μ , the distribution of X , is uniquely determined by the values it gives to half spaces. Indeed, if we know the measure of each half-space, then we know the measure of each set of the form $\{x : t \cdot x \leq \alpha\}$ (over all $t \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$). This means that we know the distribution of the variable $t \cdot X$ for all $t \in \mathbb{R}^d$ and thus the value of $\phi_{t \cdot X}(s)$ for all $t \in \mathbb{R}^d$ and $s \in \mathbb{R}$. Taking $s = 1$, we obtain the value of $\mathbb{E} e^{it \cdot X}$ for all $t \in \mathbb{R}^d$; that is, the characteristic function of X . By uniqueness we recover μ .

Billingsley remarks that this result on half-spaces is (as of the publication of the book) does not have an elementary proof without using characteristic functions.

3. There is a continuity theorem for d -dimensions: we have

$$\mu_n \rightarrow \mu \text{ weakly if and only if } \phi_n(t) \rightarrow \phi(t) \text{ point wise .}$$

The proof reduces to one-dimension using the Cramér-Wold device. See Billingsley, p.383.

4. We can also use these ideas to show that generally it suffices to reduce to the one-dimensional case.

Proposition 2.2. *Let $X_n = (X_{1,n}, \dots, X_{d,n})$ and $Y = (Y_1, \dots, Y_d)$ be random vectors. Then $X_n \Rightarrow Y$ if and only if $X_n \cdot t \Rightarrow Y \cdot t$ for all $t \in \mathbb{R}^d$.*

Proof. If $X_n \Rightarrow Y$ then since $x \mapsto t \cdot x$ is continuous, the mapping theorem implies $t \cdot X_n \Rightarrow t \cdot Y$. Conversely, if $t \cdot X_n \Rightarrow t \cdot Y$ for all $t \in \mathbb{R}^d$ then by the one-dimensional continuity theorem,

$$\mathbb{E} e^{is(t \cdot X_n)} \rightarrow \mathbb{E} e^{is(t \cdot Y)} \text{ for all } s \in \mathbb{R}, t \in \mathbb{R}^d .$$

Taking $s = 1$ gives $\phi_{X_n} \rightarrow \phi_Y$ point wise and the d -dimensional continuity theorem shows $X_n \Rightarrow Y$. $\square$

3 Gaussians and CLT

Recall that a centered Gaussian variable is one with density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}},$$

where σ^2 is the variance of the variable. The characteristic function is $\phi(t) = e^{-\frac{t^2\sigma^2}{2}}$. One natural generalization of this distribution to higher dimensions is:

Definition 3.1. A random vector $X = (X_1, \dots, X_d)$ has a (centered, or mean zero) Gaussian distribution if its characteristic function has the form

$$\phi_X(t) = \exp\left(-\frac{(tA) \cdot (tA)}{2}\right)$$

for some $d \times d$ matrix A . Here we are viewing t as a row vector.

We can ask: are there any such vectors? Well we can begin with a natural construction, given last semester in the homework. Let $X_1, \dots, X_d$ be a sequence of i.i.d. (mean zero) standard Gaussians. Forming the vector $X = (X_1, \dots, X_d)$ and calling it a *d-dimensional standard Gaussian*, we can find its characteristic function using independence and the one-dimensional formula:

$$\mathbb{E}e^{it \cdot X} = \mathbb{E}e^{i(t_1 X_1 + \dots + t_d X_d)} = \prod_{j=1}^d \mathbb{E}e^{it_j X_j} = \exp\left(-\frac{t \cdot t}{2}\right).$$

This is a *d*-dimensional standard Gaussian with matrix A equal to the identity matrix.

Generally, we can construct *d*-dimensional Gaussians by taking a linear transformation of the standard one.

Proposition 3.2. Let A be a $d \times d$ matrix and X a standard *d*-dimensional Gaussian.

1. The vector $Y = AX$ is a *d*-dimensional Gaussian associated to matrix A .
2. Defining Σ , the Covariance Matrix of Y as

$$\Sigma_{i,j} = \mathbb{E}Y_i Y_j$$

then $\Sigma = AA^T$.

3. If Y and Y' are *d*-dimensional Gaussians with the same covariance matrix, they have the same distribution: the characteristic function is given by

$$\phi_Y(t) = \exp\left(-\frac{t\Sigma t^T}{2}\right).$$

Proof. Write the entries $A_{j,k}$ of A and $X = (X_1, \dots, X_d)$. Then

$$Y_j = \sum_{k=1}^d A_{j,k} X_k .$$

We can compute the characteristic function of Y directly using independence:

$$\begin{aligned} \mathbb{E} e^{it \cdot Y} &= \mathbb{E} \exp \left(i \sum_{j=1}^d t_j \left[\sum_{k=1}^d A_{j,k} X_k \right] \right) = \mathbb{E} \exp \left(i \sum_{k=1}^d X_k \left[\sum_{j=1}^d t_j A_{j,k} \right] \right) \\ &= \prod_{k=1}^d \mathbb{E} \exp (i X_k (tA)_k) \\ &= \exp \left(-\frac{(tA) \cdot (tA)}{2} \right) . \end{aligned}$$

The covariance can be computed, again using independence: $\Sigma_{i,j}$ equals

$$\mathbb{E} Y_i Y_j = \mathbb{E} \left(\sum_{k=1}^d A_{i,k} X_k \sum_{l=1}^d A_{j,l} X_l \right) = \sum_{k=1}^d \sum_{l=1}^d A_{i,k} A_{j,l} \mathbb{E} X_k X_l = \sum_{k=1}^d A_{i,k} A_{j,k} = (AA^T)_{i,j} .$$

This proves parts 1 and 2. For 3, note that the covariance matrix determines the characteristic function, since we can write

$$\exp \left(-\frac{(tA) \cdot (tA)}{2} \right) = \exp \left(-\frac{tAA^T t^T}{2} \right) = \exp \left(-\frac{t\Sigma t^T}{2} \right) .$$

□

LECTURE 3

- Note that Σ above is positive semi-definite. This follows because it is of the form AA^T for a $d \times d$ matrix A : for all $x \in \mathbb{R}^d$,

$$\Sigma x \cdot x = A^T x \cdot A^T x \geq 0 .$$

- Conversely, if Σ is an arbitrary positive semi-definite matrix, we can factorize it as AA^T for some matrix A . In other words, we can redefine a Gaussian as one with characteristic function

$$\phi(t) = \exp \left(-\frac{t^T \Sigma t}{2} \right)$$

for some positive semi-definite matrix Σ .

3.1 CLT

There are multidimensional generalizations of the Gaussian CLT and we will state the simplest here. In the one-dimensional CLT we assumed mean zero and finite variance. In the d -dimensional case we will assume mean zero and well-defined covariance matrix. Note that if $X = (X_1, \dots, X_d)$ is a random vector then

$$|\mathbb{E}X_i X_j| \leq \|X_i\|_2 \|X_j\|_2 ,$$

so if $\mathbb{E}X_1^2 + \dots + X_d^2 < \infty$ then X has a well-defined covariance matrix.

Theorem 3.3 (d -dimensional CLT). *Let $X^{(1)}, X^{(2)}, \dots$ be an i.i.d. sequence of d -dimensional random vectors with mean zero and $\mathbb{E} \left[(X_1^{(1)})^2 + \dots + (X_d^{(1)})^2 \right] < \infty$. Letting Σ be the covariance matrix for $X^{(1)}$,*

$$\frac{X^{(1)} + \dots + X^{(n)}}{\sqrt{n}} \Rightarrow N(0, \Sigma) ,$$

where the distribution on the right is centered Gaussian with covariance matrix Σ .

Proof. Use the Cramér-Wold device: let $t \in \mathbb{R}^d$. We need only show that if Y is a $N(0, \Sigma)$ vector, then

$$t \cdot \frac{X^{(1)} + \dots + X^{(n)}}{\sqrt{n}} \Rightarrow t \cdot Y .$$

The left side is

$$\frac{X^{(1)} \cdot t + \dots + X^{(n)} \cdot t}{\sqrt{n}} ,$$

which is a normalized sum of mean-zero i.i.d. random variables with finite variance. Indeed, the mean is

$$\mathbb{E}X^{(1)} \cdot t = \sum_{i=1}^d t_i \mathbb{E}X_i^{(1)} = 0$$

and the variance is

$$\text{Var} [X^{(1)} \cdot t] = \text{Var} \sum_{i=1}^d t_i X_i^{(1)} = \sum_{i=1}^d \sum_{j=1}^d t_i t_j \mathbb{E}X_i^{(1)} X_j^{(1)} = \sum_{i=1}^d \sum_{j=1}^d t_i t_j \Sigma_{i,j} =: \sigma_t^2 .$$

So the one-dimensional CLT says this sum converges in distribution to a $N(0, \sigma_t^2)$ variable. We just need to check this is the variance of $t \cdot Y$, and this follows by exactly the same computation. □

4 Back to 1D: infinitely divisible distributions

4.1 Triangular arrays

At the end of last semester, we investigated different generalizations of the standard one-dimensional central limit theorem. The main one was due to Lindeberg, and was stated for triangular arrays. Let $(X_{n,k} : n \geq 1, k = 1, \dots, k_n)$ be an array of variables such that each row is independent:

$$X_{n,1}, \dots, X_{n,k_n} \text{ are independent for each } n .$$

We assumed that all $X_{n,k}$'s have mean zero and finite variance. A main observation was that for us to expect a central limit theorem of the form $S_n/s_n \Rightarrow N(0,1)$, where

$$S_n = \sum_{k=1}^{k_n} X_{n,k} \text{ and } s_n^2 = \text{Var } S_n ,$$

we would at least like the contribution of $\text{Var } X_{n,k}$ to the variance s_n^2 to go to zero as $n \rightarrow \infty$:

$$\frac{\max\{\sigma_{n,1}^2, \dots, \sigma_{n,k_n}^2\}}{s_n^2} \rightarrow 0 , \quad (2)$$

where $\sigma_{n,k}^2 = \text{Var } X_{n,k}$. A way to interpret this is that no variable in the n -th row contributes too much to the sum (at least measured in L^2 norm). By normalizing, we will assume

$$\sup_n s_n^2 < \infty \quad (3)$$

so that (2) implies that

$$\max\{\sigma_{n,1}^2, \dots, \sigma_{n,k_n}^2\} \rightarrow 0 . \quad (4)$$

However (2) was not quite enough for a CLT. Now we will look at what other possible limits we can get under condition (2). The first step, as in the proof of Lindeberg is

Theorem 4.1 (Accompanying Laws (simplified version)). *If (2) holds then for each $t \in \mathbb{R}$,*

$$\left| \prod_{k=1}^{k_n} \phi_{n,k}(t) - \prod_{k=1}^{k_n} e^{\phi_{n,k}(t)-1} \right| \rightarrow 0 \text{ as } n \rightarrow \infty .$$

Here, $\phi_{n,k}(t)$ is the characteristic function of $X_{n,k}$.

What is the meaning of this theorem? Let's discuss this before the proof.

1. The product $\prod_{k=1}^{k_n} \phi_{n,k}(t)$ is the characteristic function of S_n . The other is the characteristic function of a somewhat weirder variable. To see what it is, consider replacing each summand $X_{n,k}$ in the sum S_n by a random number of copies of itself. In other

words, let $N_1, \dots, N_{k_n}$ be i.i.d. random variables with Poisson distribution (with parameter 1). Now replace each $X_{n,k}$ by a sum of N_k i.i.d. copies of itself: define

$$Y_{n,k} = X_{n,k}^{(1)} + \dots + X_{n,k}^{(N_k)},$$

where the summands on the right are i.i.d. (One way to think of this is the value at time 1 of a compound Poisson process – this was in last semester’s final.) Then you can check that the characteristic function of $Y_{n,k}$ is $e^{\phi_{n,k}(t)-1}$. Therefore if we define

$$T_n = Y_{n,1} + \dots + Y_{n,k_n}$$

then the characteristic function of T_n is $\prod_{k=1}^{k_n} e^{\phi_{n,k}(t)-1}$.

2. The real theorem on accompanying laws actually holds without the variance assumption (2), but with a weaker, but similar, condition that all variables in the n -th row are small. The theorem is slightly different, but asserts that we can replace all sums with “Poissonized” versions. See, for instance, Varadhan’s book. We will, however, assume (2).

LECTURE 4

Recall we are proving the following theorem on triangular arrays $(X_{n,k} : n \geq 1, k = 1, \dots, k_n)$ that satisfy the conditions

$$\sup_n s_n^2 < \infty \tag{5}$$

and

$$\max\{\sigma_{n,1}^2, \dots, \sigma_{n,k_n}^2\} \rightarrow 0. \tag{6}$$

Theorem 4.2 (Accompanying Laws (simplified version)). *If (12) and (6) hold then for each $t \in \mathbb{R}$,*

$$\left| \prod_{k=1}^{k_n} \phi_{n,k}(t) - \prod_{k=1}^{k_n} e^{\phi_{n,k}(t)-1} \right| \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Here, $\phi_{n,k}(t)$ is the characteristic function of $X_{n,k}$.

Proof. We will use a result that got a good amount of traction last semester. Recall that it was proved using induction.

Lemma 4.3. *If $w_1, \dots, w_n$ and $z_1, \dots, z_n$ are complex numbers of absolute value ≤ 1 then*

$$|w_1 \cdots w_n - z_1 \cdots z_n| \leq \sum_{i=1}^n |w_i - z_i|.$$

Fix $t \in \mathbb{R}$. We must check that $w_i = \phi_{n,i}(t)$ and $z_i = e^{\phi_{n,i}(t)-1}$ satisfy the hypotheses of the lemma. However they are both characteristic functions (z_i is the characteristic function of a compound Poisson process at time 1). So we can estimate

$$\left| \prod_{k=1}^{k_n} \phi_{n,k}(t) - \prod_{k=1}^{k_n} e^{\phi_{n,k}(t)-1} \right| \leq \sum_{k=1}^{k_n} |\phi_{n,k}(t) - e^{\phi_{n,k}(t)-1}| .$$

Set $\theta_{n,k} = \phi_{n,k}(t) - 1$. Then this equals

$$\sum_{k=1}^{k_n} |e^{\theta_{n,k}} - 1 - \theta_{n,k}| .$$

However we can estimate

$$|e^z - 1 - z| \leq \sum_{n=2}^{\infty} \frac{|z|^n}{n!} = |z|^2 \sum_{n=0}^{\infty} \frac{|z|^n}{(n+2)!} \leq |z|^2 e^{|z|} ,$$

so the above is bounded by

$$\sum_{k=1}^{k_n} |\theta_{n,k}|^2 e^{|\theta_{n,k}|} .$$

Next we should estimate $\theta_{n,k}$ and use the mean-zero assumption:

$$|\theta_{n,k}| \leq \mathbb{E} |e^{itX_{n,k}} - 1 - itX_{n,k}| .$$

By the general inequality

$$\left| e^{ix} - \sum_{k=0}^n \frac{(ix)^k}{k!} \right| \leq \min \left\{ \frac{2|x|^n}{n!}, \frac{|x|^{n+1}}{(n+1)!} \right\}$$

(proved in Billingsley (26.1)-(26.4) for $n = 1$, we get the bound

$$|\theta_{n,k}| \leq \mathbb{E} X_{n,k}^2 / 2 = \sigma_{n,k}^2 / 2 .$$

This implies by (12) that

$$\sum_{k=1}^{k_n} |\theta_{n,k}| \text{ is bounded as } n \rightarrow \infty .$$

Now we use (6) to deduce

$$\sum_{k=1}^{k_n} |\theta_{n,k}|^2 e^{|\theta_{n,k}|} \leq \max_{k=1, \dots, k_n} |\theta_{n,k}| e^{|\theta_{n,k}|} \sum_{k=1}^{k_n} |\theta_{n,k}| \rightarrow 0$$

and we are done. □

The advantage of using the Poissonized variables, that is, using the product $\prod_{k=1}^{k_n} e^{\phi_{n,k}(t)-1}$, is that we can get a nice representation for them:

$$\exp \left(\sum_{k=1}^{k_n} \phi_{n,k}(t) - 1 \right) = \exp \left(\sum_{k=1}^{k_n} \int e^{itx} - 1 \, d\mu_{n,k}(x) \right) ,$$

where $\mu_{n,k}$ is the distribution of $X_{n,k}$. We can further reduce this by using the fact that $\mathbb{E}X_{n,k} = 0$, so this equals

$$\exp \left(\sum_{k=1}^{k_n} \int e^{itx} - 1 - itx \, d\mu_{n,k}(x) \right) = \exp \left(\sum_{k=1}^{k_n} \int \frac{e^{itx} - 1 - itx}{x^2} x^2 \, d\mu_{n,k}(x) \right) .$$

Dividing by x^2 is ok here since we know that

$$\frac{e^{itx} - 1 - itx}{x^2} \rightarrow -\frac{t^2}{2} \text{ as } x \rightarrow 0 .$$

If we define the measure μ_n by $\mu_n = \sum_{k=1}^{k_n} \mu_{n,k}$ then we obtain

$$\exp \left(\int \frac{e^{itx} - 1 - itx}{x^2} x^2 \, d\mu_n(x) \right) ,$$

and finally if we define ν_n by

$$\nu_n((-\infty, y]) = \int_{-\infty}^y x^2 \mu_n(x) ,$$

then the above equals

$$\exp \left(\int \frac{e^{itx} - 1 - itx}{x^2} \, d\nu_n(x) \right) .$$

This is the representation we are going for.

Theorem 4.4 (Limit theorem for triangular arrays). *Let $(X_{n,k} : n \geq 1, k = 1, \dots, k_n)$ be an array with each row independent, satisfying (12) and (6). Setting $S_n = X_{n,1} + \dots + X_{n,k_n}$, if $S_n \Rightarrow Y$ for some random variable Y then*

$$\phi_Y(t) = \exp \left(\int \frac{e^{itx} - 1 - itx}{x^2} \, d\mu(x) \right) \text{ for all } t \in \mathbb{R}$$

for some finite measure μ on $\mathbb{R}$.

Proof. The accompanying laws theorem implies that if we set S'_n to be a variable with characteristic function

$$\prod_{k=1}^{k_n} e^{\phi_{n,k}-1} = \exp \left(\int \frac{e^{itx} - 1 - itx}{x^2} \, d\nu_n(x) \right) ,$$

then $S'_n \Rightarrow Y$. To get a representation for Y we will try to simply take a limit of the measures ν_n . Unfortunately these are not probability measures, so we cannot directly apply the theory of weak convergence. However, for finite measures there is a related concept, *vague convergence*, which operates similarly. We will now summarize the relevant results.

A sequence (ν_n) of finite measures converges vaguely to a measure ν if for each continuous $f : \mathbb{R} \rightarrow \mathbb{R}$ with $\lim_{|x| \rightarrow \infty} f(x) = 0$, one has

$$\int f \, d\nu_n \rightarrow \int f \, d\nu .$$

This can be seen as a generalization of weak convergence in the case of probability measures. However the condition on the large- x behavior of f essentially allows for cases in which mass escapes to infinity. For example, we can have a sequence of probability measures converging vaguely to the zero measure. For this reason we do not need tightness for convergent subsequences: if we know simply that

$$\sup_n \nu_n(\mathbb{R}) < \infty$$

then we can find a measure ν and a subsequence (ν_{n_k}) such that $\nu_{n_k} \rightarrow \nu$ vaguely.

By (12),

$$\sup_n \nu_n(\mathbb{R}) = \sup_n \sum_{k=1}^{k_n} \int x^2 \, d\mu_{n,k}(x) = \sup_n s_n^2 < \infty .$$

Therefore we can find a subsequence (ν_{n_k}) and a measure μ such that $\nu_{n_k} \rightarrow \mu$ vaguely. Since

$$\lim_{|x| \rightarrow \infty} \frac{e^{itx} - 1 - itx}{x^2} = 0 \text{ for each } t ,$$

we also have

$$\prod_{k=1}^{k_{n_k}} e^{\phi_{n_k,k}(t)-1} \rightarrow \exp \left(\int \frac{e^{itx} - 1 - itx}{x^2} \, d\mu(x) \right)$$

and therefore

$$\prod_{k=1}^{k_{n_k}} \phi_{n_k,k}(t) \rightarrow \exp \left(\int \frac{e^{itx} - 1 - itx}{x^2} \, d\mu(x) \right) .$$

The left side is the characteristic function of S_{n_k} , so since $S_{n_k} \Rightarrow Y$, the right side is the characteristic function of Y . $\square$

5 Infinitely divisible distributions

We will see in this section that there is a different classification of the limits from the previous section: they are the infinitely divisible distributions. First let us understand a little better the functions from last section: for a finite measure μ on $\mathbb{R}$, define

$$\Phi(\mu, t) = \exp \left(\int \frac{e^{itx} - 1 - itx}{x^2} d\mu(x) \right) .$$

The integrand is defined as $-t^2/2$ at $x = 0$.

Examples.

1. The characteristic function of a Gaussian with mean 0 and variance σ^2 is

$$\phi(t) = e^{itm} e^{-\frac{t^2 \sigma^2}{2}} = \exp \left(\int \frac{e^{itx} - 1 - itx}{x^2} d\mu(x) \right) ,$$

where μ is a delta mass of size σ^2 at 0: $\mu = \sigma^2 \delta_0$. Here the Gaussian has mean zero and variance $\mu(\mathbb{R})$.

2. If X is Poisson with mean m then its characteristic function is

$$\mathbb{E} e^{itX} = e^{-m} \sum_{k=0}^{\infty} \frac{(e^{it} m)^k}{k!} = e^{m(e^{it} - 1)} .$$

Therefore the characteristic function of $\lambda(X - m)$, where $\lambda > 0$, is

$$\begin{aligned} \mathbb{E} e^{i(t\lambda)X} e^{-it\lambda m} &= e^{m(e^{it\lambda} - 1)} e^{-it\lambda m} = \exp \left(m(e^{it\lambda} - it\lambda - 1) \right) \\ &= \exp \left(\lambda^2 m \frac{e^{it\lambda} - it\lambda - 1}{\lambda^2} \right) , \end{aligned}$$

and this equals $\Phi(\lambda^2 m \delta_\lambda, t)$. In other words, $\lambda(X - m)$ corresponds to the measure $\lambda^2 m \delta_\lambda$.

3. Said differently, the point mass $a\delta_b$ corresponds to the distribution of the variable $b(X - (a/b^2))$, where X is Poisson of mean a/b^2 . Therefore the mean of $b(X - (a/b^2))$ is zero and the variance is a .
4. If μ_1 and μ_2 are finite measures then $\Phi(\mu_1 + \mu_2, t) = \Phi(\mu_1, t)\Phi(\mu_2, t)$.

Theorem 5.1. *For each finite measure μ on $\mathbb{R}$, $\Phi(\mu, t)$ is the characteristic function of a variable with mean zero and variance $\mu(\mathbb{R})$.*

Proof. If μ is a point mass then we have already proved $\Phi(\mu, t)$ is a characteristic function of a variable with mean zero and variance $\mu(\mathbb{R})$. For any sum of point masses $\mu = \sum_{i=1}^n a_i \delta_{x_i}$, we have

$$\Phi(\mu, t) = \prod_{i=1}^n \Phi(a_i \delta_{x_i}, t) ,$$

so it is the characteristic function of a sum of independent variables with mean zero and variance a_i . This sum thus has mean zero and variance $\sum_i a_i = (\sum_i a_i \delta_{x_i})(\mathbb{R})$.

In the homework you will show that any finite measure μ can be written as a vague limit of a sequence (μ_n) such that each μ_n is a finite sum of point masses and $\sup_n \mu_n(\mathbb{R}) < \infty$. So given a finite measure μ , choose a sequence (μ_n) of finite measures (which are finite sums of point masses) that converges vaguely to μ . Then $\Phi(\mu_n, t) \rightarrow \Phi(\mu, t)$ point wise (in t) and since each $\Phi(\mu_n, t)$ is a characteristic function, we need only show that $\Phi(\mu, t)$ is continuous at zero to show that it is a characteristic function. (We proved this last semester.) So we must show that if $t_n \rightarrow 0$ then $\Phi(\mu, t_n) \rightarrow \Phi(\mu, 0)$, or

$$\exp \left(\int \frac{e^{it_n x} - 1 - it_n x}{x^2} d\mu(x) \right) \rightarrow 1 .$$

However by the bound

$$|e^{ix} - 1 - ix| \leq \frac{|x|^2}{2} ,$$

we obtain

$$\left| \int \frac{e^{it_n x} - 1 - it_n x}{x^2} d\mu(x) \right| \leq \frac{|t_n|^2}{2} \mu(\mathbb{R}) \rightarrow 0 ,$$

so $\Phi(\mu, t)$ is a characteristic function.

This means that if X_n is a variable with characteristic function $\Phi(\mu_n, t)$ then $X_n \Rightarrow X$, where X has characteristic function $\Phi(\mu, t)$. We can compute the mean and variance of X as follows. First the second moment is finite by Fatou:

$$\mathbb{E}X^2 \leq \liminf_n \mathbb{E}X_n^2 = \liminf_n \mu_n(\mathbb{R}) < \infty .$$

Since X has two moments, we can compute them by taking derivatives of the characteristic function:

$$\begin{aligned} \frac{d}{dt} \exp \left(\int \frac{e^{itx} - 1 - itx}{x^2} d\mu(x) \right) \Big|_{t=0} &= \frac{d}{dt} \int \frac{e^{itx} - 1 - itx}{x^2} d\mu(x) \Big|_{t=0} \\ &= i \int \frac{e^{itx} - 1}{x} d\mu(x) \Big|_{t=0} = 0 . \end{aligned}$$

(Some care needs to be taken with the above computation, in particular being able to pass the derivative through the integral. You can check this.) Similarly, we can compute

$$\frac{d}{dt} i \int \frac{e^{itx} - 1}{x} d\mu(x) \Big|_{t=0} = - \int e^{itx} d\mu(x) \Big|_{t=0} = -\mu(\mathbb{R}) .$$

This means $\text{Var } Y = \mu(\mathbb{R})$. □

Not only is $\Phi(\mu, t)$ a characteristic function, the associated distribution is infinitely divisible:

Definition 5.2. *A distribution μ is infinitely divisible if for each $n \geq 1$, we can find a distribution μ_n such that*

$$\mu = \mu_n * \cdots * \mu_n ,$$

where the right side is a convolution of n copies of μ_n .

Equivalently, a distribution μ of a random variable X is infinitely divisible if for each n , we can decompose X as a sum of n i.i.d. copies of another variable X_n . Equivalently, if the characteristic function of μ is ϕ then we can find another characteristic function ϕ_n such that $\phi = \phi_n^n$.

Theorem 5.3 (Characterization theorem). *Y is infinitely divisible with mean zero and finite variance if and only if there is a finite measure μ such that*

$$\phi_Y(t) = \Phi(\mu, t) .$$

Proof. Suppose that $\phi_Y(t) = \Phi(\mu, t)$. Then for any n , $\Phi(\mu, t) = \Phi(\mu/n, t)^n$, so since all of these are characteristic functions, the distribution of Y is infinitely divisible and Y has mean zero and variance $\mu(\mathbb{R}) < \infty$.

Next time we will prove the converse. □

LECTURE 6

Last time we proved one direction of:

Theorem 5.4 (Characterization theorem). *Y is infinitely divisible with mean zero and finite variance if and only if there is a finite measure μ such that*

$$\phi_Y(t) = \Phi(\mu, t) .$$

Proof. Now we prove the other direction. Conversely suppose that Y is infinitely divisible with mean zero and finite variance. Then for each n , let $X_{n,1}, \dots, X_{n,n}$ be an i.i.d. sequence such that $X_{n,1} + \cdots + X_{n,n}$ has the same distribution as Y . The array $(X_{n,k} : n \geq 1, k = 1, \dots, n)$ then satisfies the conditions for the triangular array limit theorem:

1. Since Y has two moments, $X_{n,1} + \cdots + X_{n,n}$ has two moments. We deduce from this that each $X_{n,k}$ has two moments. To do this, we prove that if X and Z are independent and $X + Z$ has two moments, then so does X and Z .

Proof. $X^2 + Z^2 \leq (X + Z)^2 + 2|XZ|$, so since $(X + Z)^2$ is integrable, we need only show that $|XZ|$ is integrable. By Fubini, we can show that $|X|$ and $|Z|$ are integrable. If $|X|$ were not integrable, then because $|X| \leq |X + z| + |z|$, we would have $|X + z|$ not integrable for each z . This, by Fubini, would imply that $|X + Z|$ is not integrable, which contradicts $(X + Z)^2$ integrable. □

2. Each $X_{n,k}$ has two moments, so it has one moment. By the i.i.d. assumption, we have $n\mathbb{E}X_{n,k} = \mathbb{E}Y = 0$, so $\mathbb{E}X_{n,k} = 0$. Furthermore,

$$\frac{\max\{\sigma_{n,1}^2, \dots, \sigma_{n,n}^2\}}{s_n^2} = \frac{\text{Var } X_{n,1}}{\text{Var } Y} = 1/n \rightarrow 0 .$$

Finally,

$$\sup_n s_n^2 = \sup_n \mathbb{E}Y^2 = \text{Var } Y < \infty .$$

Since $S_n \Rightarrow Y$ (in fact, all the row sums S_n have the same distribution – that of Y), Y has characteristic function $\Phi(\mu, t)$ for some finite measure μ and is thus infinitely divisible. $\square$

To finish, let us state another theorem about the representation (without proof). It can be useful to find the measure μ in the characteristic function $\Phi(\mu, t)$. The reason is:

Theorem 5.5. *Let $(X_n), X$ be infinitely divisible with mean zero and finite variance and characteristic functions $\Phi(\mu_n, t), \Phi(\mu, t)$. Then $X_n \Rightarrow X$ if and only if μ_n converges to μ vaguely.*

Proof. See the proof in Billingsley, Theorem 28.4. $\square$

- In the case that X is normal, we have $\mu = \delta_0$. So if we have a triangular array satisfying the conditions from the previous lectures:

$$\sup_n s_n^2 < \infty \text{ and } \max\{\sigma_{n,1}^2, \dots, \sigma_{n,k_n}^2\} \rightarrow 0 ,$$

the Poissonized row sum (at level n) has characteristic function $\Phi(\mu_n, t)$, where

$$\mu_n(-\infty, x] = \sum_{k=1}^{k_n} \int_{-\infty}^x y^2 \, dF_{n,k}(y) .$$

When $s_n^2 = 1$ for all n , we have $\mu_n \rightarrow \delta_0$ vaguely if and only if

$$\sum_{k=1}^{k_n} \int_{-\epsilon}^{\epsilon} y^2 \, dF_{n,k}(y) \rightarrow 0 ,$$

or

$$\sum_{k=1}^{k_n} \mathbb{E}X_{n,k} \mathbf{1}_{\{|X_{n,k}| \geq \epsilon\}} \rightarrow 0 ,$$

which is Lindeberg.

6 Derivatives and conditional probability

6.1 Conditional expectation

We will begin by talking about conditional expectation. Recall the “basic” definition of conditional probability from last semester: if $\mathbb{P}(A) > 0$ then the probability $\mathbb{P}(B \mid A)$ is defined as

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} .$$

We can make a similar definition of conditional expectation. If X is an integrable random variable with $\mathbb{P}(A) > 0$ then we can give a “basic” definition of conditional expectation as

$$\mathbb{E}(X \mid A) = \frac{\mathbb{E}X\mathbf{1}_A}{\mathbb{P}(A)} .$$

Note that in the case $X = \mathbf{1}_B$ this definition reduces to that of the conditional probability $\mathbb{P}(B \mid A)$.

Now we will give the real definition of conditional expectation, and from this we will define conditional probability. As you can see above, one main problem with the basic definition is that it makes no sense if $\mathbb{P}(A) = 0$. We will give definitions which essentially deal with this problem, and also quite greatly expand the scope of the concept.

The first difference is that, although the basic definition of condition expectation gives a number, our real definition will give a function. To get an idea of how this could happen, consider an intermediate case: if $0 < \mathbb{P}(A) < 1$ and X is integrable, we can define a function $f : \Omega \rightarrow \mathbb{R}$ by

$$f(\omega) = \begin{cases} \frac{\mathbb{E}X\mathbf{1}_A}{\mathbb{P}(A)} & \text{if } \omega \in A \\ \frac{\mathbb{E}X\mathbf{1}_{A^c}}{\mathbb{P}(A^c)} & \text{if } \omega \in A^c \end{cases} .$$

This is a step function which takes the value from the basic definition on each piece A and A^c of the space.

The second difference is that we will not condition on events any more; we will condition on sigma-algebras. The function f defined above corresponds to conditioning on the sigma-algebra generated by the event A . In other words, we are assuming that the observer of our experiment (whose outcome is modeled by ω) knows all information contained in this sigma-algebra, but no more: they know if $\omega \in A$ or if $\omega \in A^c$. Then the observer can determine the expectation of X given their information. To condition on general sigma-algebras is considerably more involved, but the method will also allow to deal with sets of measure zero.

So my advice is to first absorb the definitions, learn their properties, and how to manipulate them. Later on, we will relate our definition to the basic ones.

Definition 6.1. *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space with X an integrable random variable and $\mathcal{F}$ a sub-sigma-algebra of Σ . The conditional expectation $\mathbb{E}[X \mid \mathcal{F}]$ is the unique (up to measure zero) random variable such that*

1. $\mathbb{E}[X \mid \mathcal{F}]$ is measurable relative to $\mathcal{F}$ and

2. for all events $A \in \mathcal{F}$, one has

$$\mathbb{E}[X \mathbf{1}_A] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{F}] \mathbf{1}_A] .$$

Written in terms of integrals,

$$\int_A X \, d\mathbb{P} = \int_A \mathbb{E}[X \mid \mathcal{F}] \, d\mathbb{P} . \quad (7)$$

If Y is another random variable, we define $\mathbb{E}[X \mid Y]$ to be $\mathbb{E}[X \mid \sigma(Y)]$.

Theorem 6.2 (Existence and uniqueness). *Given X integrable and $\mathcal{F}$ a sub-sigma-algebra of Σ , there exists an (almost surely) unique $\mathcal{F}$ -measurable random variable satisfying (7).*

Proof. We will apply the Radon-Nikodym theorem.

Lemma 6.3 (Radon-Nikodym). *Let μ be a finite measure on (Ω, Σ) that is absolutely continuous relative to $\mathbb{P}$; that is, for all A with $\mathbb{P}(A) = 0$, we have $\mu(A) = 0$. Then there exists a Σ -measurable, integrable (relative to $\mathbb{P}$) function $f : \Omega \rightarrow \mathbb{R}$ such that*

$$\nu(A) = \int_A f \, d\mathbb{P} \text{ for all } A \in \Sigma .$$

Furthermore for any f, g satisfying the above, we have $f = g$ $\mathbb{P}$ -almost surely.

To be continued. □

LECTURE 7

We were proving:

Theorem 6.4 (Existence and uniqueness). *Given X integrable and $\mathcal{F}$ a sub-sigma-algebra of Σ , there exists an (almost surely) unique $\mathcal{F}$ -measurable random variable satisfying*

$$\int_A \mathbb{E}[X \mid \mathcal{F}] \, d\mathbb{P} = \int_A X \, d\mathbb{P}, \quad A \in \mathcal{F} .$$

Proof. We now apply the Radon-Nikodym theorem in our context. We can define the measure $\mathbb{P}$ only on $\mathcal{F}$: that is, we set $\mathbb{P}_{\mathcal{F}} : \mathcal{F} \rightarrow [0, 1]$ by

$$\mathbb{P}_{\mathcal{F}}(A) = \mathbb{P}(A) \text{ for } A \in \mathcal{F} .$$

Furthermore, define $\nu_+ : \mathcal{F} \rightarrow [0, 1]$ by

$$\nu_+(A) = \int_A X_+ \, d\mathbb{P} .$$

Because X_+ is non-negative and integrable, ν_+ is a finite measure. Furthermore, it is absolutely continuous relative to $\mathbb{P}_{\mathcal{F}}$: if $A \in \mathcal{F}$ satisfies $\mathbb{P}_{\mathcal{F}}(A) = 0$ then $\mathbb{P}(A) = 0$ and therefore

$\nu_+(A) = \int_A X_+ \, d\mathbb{P} = 0$. So by Radon-Nikodym, we can find a function $f : \Omega \rightarrow \mathbb{R}$ that is integrable relative to $\mathbb{P}_{\mathcal{F}}$ and measurable relative to $\mathcal{F}$ such that

$$\int_A X_+ \, d\mathbb{P} = \nu_+(A) = \int_A f \, d\mathbb{P}_{\mathcal{F}} = \int f \mathbf{1}_A \, d\mathbb{P}_{\mathcal{F}} .$$

But one can check that for any $\mathcal{F}$ -measurable function Y that is integrable relative to $\mathbb{P}_{\mathcal{F}}$, one has $\int Y \, d\mathbb{P}_{\mathcal{F}} = \int Y \, d\mathbb{P}$ (begin with indicator functions, then simple functions, then take limits). This means in particular that Y is $\mathbb{P}$ -integrable. So we obtain

$$\int_A X_+ \, d\mathbb{P} = \int_A f \, d\mathbb{P} \text{ for } A \in \mathcal{F}$$

and f satisfies the conditions of the theorem with X replaced by X_+ .

We can then repeat the construction with X_- to obtain g which is integrable and measurable relative to $\mathcal{F}$ such that

$$\int_A X_- \, d\mathbb{P} = \int_A g \, d\mathbb{P} \text{ for } A \in \mathcal{F} .$$

Now setting $\mathbb{E}[X \mid \mathcal{F}] = f - g$ we are done with existence.

For uniqueness, we simply appeal to the result we proved last semester. If $h, k : \Omega \rightarrow \mathbb{R}$ are integrable such that

$$\int_A h \, d\mathbb{P} = \int_A k \, d\mathbb{P} \text{ for all } A \in \Sigma$$

then $h = k$ almost surely. Apply this to the measure $\mathbb{P}_{\mathcal{F}}$ on the sigma-algebra $\mathcal{F}$ to obtain uniqueness. $\square$

The value of $\mathbb{E}[X \mid \mathcal{F}]$ is to be interpreted as follows. Suppose an observer knows all information contained in the sub-sigma-algebra $\mathcal{F}$. Then, although your best guess for the value of X on the outcome ω is $\mathbb{E}X$ (since you have no information), the observer's best guess for the value of X on the outcome ω is $\mathbb{E}[X \mid \mathcal{F}](\omega)$.

Properties of conditional expectation. Here we give many properties. Working through them will give you an idea of the mechanics of conditional expectation. For all of these properties, let X, Y and X_n be integrable for each n and $\mathcal{F}$ a sub-sigma-algebra of Σ .

1. (Linearity) If $a, b \in \mathbb{R}$ then almost surely

$$\mathbb{E}[aX + bY \mid \mathcal{F}] = a\mathbb{E}[X \mid \mathcal{F}] + b\mathbb{E}[Y \mid \mathcal{F}] .$$

Proof. Let $A \in \mathcal{F}$. Then

$$\begin{aligned} \int_A (aX + bY) \, d\mathbb{P} &= a \int_A X \, d\mathbb{P} + b \int_A Y \, d\mathbb{P} \\ &= a \int_A \mathbb{E}[X \mid \mathcal{F}] \, d\mathbb{P} + b \int_A \mathbb{E}[Y \mid \mathcal{F}] \, d\mathbb{P} \\ &= \int_A (a\mathbb{E}[X \mid \mathcal{F}] + b\mathbb{E}[Y \mid \mathcal{F}]) \, d\mathbb{P} . \end{aligned}$$

By uniqueness of conditional expectation, we obtain the result. $\square$

2. If X is itself $\mathcal{F}$ -measurable and XY is integrable then $\mathbb{E}[XY | \mathcal{F}] = X\mathbb{E}[Y | \mathcal{F}]$. In this case, the observer learns the full value of X from their information $\mathcal{F}$.

Proof. First take $X = \mathbf{1}_B$ for some $B \in \mathcal{F}$. Take $A \in \mathcal{F}$ arbitrary and compute

$$\int_A XY \, d\mathbb{P} = \int_{A \cap B} Y \, d\mathbb{P} = \int_{A \cap B} \mathbb{E}[Y | \mathcal{F}] \, d\mathbb{P}.$$

The last equality holds because $A \cap B \in \mathcal{F}$. This equals $\int_A X\mathbb{E}[Y | \mathcal{F}] \, d\mathbb{P}$, and completes the proof in this case.

For any nonnegative simple variable $X = \sum_i a_i \mathbf{1}_{A_i}$ for $A_i \in \mathcal{F}$ we use linearity and the previous case to find

$$\mathbb{E}[XY | \mathcal{F}] = \sum_i a_i \mathbb{E}[\mathbf{1}_{A_i} Y | \mathcal{F}] = \sum_i a_i \mathbf{1}_{A_i} \mathbb{E}[Y | \mathcal{F}] = Y \mathbb{E}[X | \mathcal{F}].$$

If X is nonnegative and $\mathcal{F}$ -measurable, we can write it as a monotone limit of simple $\mathcal{F}$ -measurable functions (S_n) . For $A \in \mathcal{F}$, use monotone convergence:

$$\int_A XY_+ \, d\mathbb{P} = \int_A \lim_n S_n Y_+ \, d\mathbb{P} = \lim_n \int_A S_n \mathbb{E}[Y_+ | \mathcal{F}] \, d\mathbb{P} = \int_A X \mathbb{E}[Y_+ | \mathcal{F}] \, d\mathbb{P}.$$

This means $\mathbb{E}[XY_+ | \mathcal{F}] = X\mathbb{E}[Y_+ | \mathcal{F}]$. By linearity,

$$\mathbb{E}[XY | \mathcal{F}] = X\mathbb{E}[Y | \mathcal{F}],$$

and this completes the case $X \geq 0$. For the general case of X integrable, just split $X = X_+ - X_-$ and use linearity. $\square$

3. If $X = a$ almost surely then $\mathbb{E}[X | \mathcal{F}] = a$ almost surely.

Proof. Take $Y = 1$ in the previous property, noting that since X is constant, it is $\mathcal{F}$ -measurable. $\square$

4. If X is independent of $\mathcal{F}$ then $\mathbb{E}[X | \mathcal{F}] = \mathbb{E}X$. In particular, $\mathbb{E}[X | \{\emptyset, \Omega\}] = \mathbb{E}X$. In this case, the observer learns nothing from their information $\mathcal{F}$.

Proof. For $A \in \mathcal{F}$,

$$\int_A X \, d\mathbb{P} = \mathbb{E}[X \mathbf{1}_A] = \mathbb{E}X \mathbb{P}(A) = \mathbb{E}X \int_A d\mathbb{P} = \int_A (\mathbb{E}X) \, d\mathbb{P}.$$

$\square$

LECTURE 8

We continue with properties of conditional expectation.

1. If $X \leq Y$ almost surely then $\mathbb{E}[X \mid \mathcal{F}] \leq \mathbb{E}[Y \mid \mathcal{F}]$ almost surely.

Proof. Suppose it is false, so that the event

$$A := \{\mathbb{E}[X \mid \mathcal{F}] - \mathbb{E}[Y \mid \mathcal{F}] > 0\}$$

has positive probability. Note that $A \in \mathcal{F}$ since the variable $\mathbb{E}[X \mid \mathcal{F}] - \mathbb{E}[Y \mid \mathcal{F}]$ is $\mathcal{F}$ -measurable. Therefore

$$0 < \int_A (\mathbb{E}[X \mid \mathcal{F}] - \mathbb{E}[Y \mid \mathcal{F}]) \, d\mathbb{P}.$$

By linearity of conditional expectation, the integrand is $\mathbb{E}[X - Y \mid \mathcal{F}]$, so we obtain

$$0 < \int_A (X - Y) \, d\mathbb{P}.$$

But $X \leq Y$ almost surely, so $(X - Y)\mathbf{1}_A \leq 0$ almost surely and this is a contradiction. $\square$

2. $|\mathbb{E}[X \mid \mathcal{F}]| \leq \mathbb{E}[|X| \mid \mathcal{F}]$.

Proof. This follows since $X \leq |X|$ and $-X \leq |X|$. Then apply the last property. $\square$

3. (Dominated convergence) If $X_n \rightarrow X$ almost surely, $|X_n| \leq Y$ for all n , then $\mathbb{E}[X_n \mid \mathcal{F}] \rightarrow \mathbb{E}[X \mid \mathcal{F}]$ almost surely.

Proof. Setting $Z_n = \mathbb{E}[X_n \mid \mathcal{F}]$ and $Z = \mathbb{E}[X \mid \mathcal{F}]$, we must show that $Z_n \rightarrow Z$ almost surely. Assume first that $X = 0$ and that X_n decreases point wise to 0. In this case, we must show that $Z_n \rightarrow 0$ almost surely. By the previous monotonicity property, $Z_{n+1} \leq Z_n$ almost surely for all n so if Z_n does not converge to 0, we can find $\epsilon > 0$ such that

$$\mathbb{P}(Z_n > \epsilon \text{ for all } n) > \epsilon.$$

Calling this event A , it is $\mathcal{F}$ -measurable and so using dominated convergence,

$$\int_A Z_n \, d\mathbb{P} = \int \mathbf{1}_A X_n \, d\mathbb{P} \rightarrow 0,$$

a contradiction since $\int_A Z_n \, d\mathbb{P} > \epsilon^2$ for all n . Thus $Z_n \rightarrow 0$ almost surely.

To deal with the general case, define

$$U_n = \sup_{k \geq n} |X_k - X|,$$

so that U_n decreases to 0 almost surely. Since $U_n \leq 2Y$ we can use the previous case to say that $\mathbb{E}[U_n \mid \mathcal{F}] \rightarrow 0$ almost surely. But now

$$|\mathbb{E}[X_n \mid \mathcal{F}] - \mathbb{E}[X \mid \mathcal{F}]| \leq \mathbb{E}[|X_n - X| \mid \mathcal{F}] \leq \mathbb{E}[U_n \mid \mathcal{F}] \rightarrow 0$$

almost surely and we are done. $\square$

4. (Jensen's inequality) Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be convex and assume $\phi(X)$ is integrable. Then

$$\phi(\mathbb{E}[X \mid \mathcal{F}]) \leq \mathbb{E}[\phi(X) \mid \mathcal{F}] \text{ almost surely .}$$

Proof. The proof is similar to the unconditional version and uses the fact that for each $x_0 \in \mathbb{R}$ we can find a number a such that

$$\phi(x_0) + (x - x_0)a \leq \phi(x) \text{ for all } x \in \mathbb{R} .$$

This is a rephrasing of the fact that there is a supporting line for the graph of ϕ at each point. We may choose $a = a(x_0)$ to be the right derivative of ϕ at x_0 . So with this choice, taking $x_0 = \mathbb{E}[X \mid \mathcal{F}]$, we get for almost every ω ,

$$\phi(\mathbb{E}[X \mid \mathcal{F}]) + (X - \mathbb{E}[X \mid \mathcal{F}])a(\mathbb{E}[X \mid \mathcal{F}]) \leq \phi(X) .$$

Assuming first that $\mathbb{E}[X \mid \mathcal{F}]$ is almost surely bounded, then all the terms above are integrable and we can take expected value conditioned on $\mathcal{F}$ on both sides. The first term is, using $\mathcal{F}$ -measurability,

$$\mathbb{E}[\phi(\mathbb{E}[X \mid \mathcal{F}]) \mid \mathcal{F}] = \phi(\mathbb{E}[X \mid \mathcal{F}]) .$$

The second is

$$\mathbb{E}[(X - \mathbb{E}[X \mid \mathcal{F}])\mathbb{E}[X \mid \mathcal{F}] \mid \mathcal{F}] = 0$$

and the third is $\mathbb{E}[\phi(X) \mid \mathcal{F}]$. We then obtain the statement of Jensen's inequality.

For unbounded $\mathbb{E}[X \mid \mathcal{F}]$, we just approximate, setting

$$G_n = \{|\mathbb{E}[X \mid \mathcal{F}]| \leq n\} .$$

$X\mathbf{1}_{G_n}$ satisfies the conditions of the previous case, so we can apply it to obtain

$$\phi(\mathbb{E}[X\mathbf{1}_{G_n} \mid \mathcal{F}]) \leq \mathbb{E}[\phi(X\mathbf{1}_{G_n}) \mid \mathcal{F}] .$$

The left side equals $\phi(\mathbf{1}_{G_n}\mathbb{E}[X \mid \mathcal{F}])$ and converges point wise to $\phi(\mathbb{E}[X \mid \mathcal{F}])$. For the right side, since X is integrable and $\mathbf{1}_{G_n} \rightarrow 1$,

$$X\mathbf{1}_{G_n} \rightarrow X, \text{ so } \phi(X\mathbf{1}_{G_n}) \rightarrow \phi(X) \text{ almost surely .}$$

Again, since X is integrable and dominates $X\mathbf{1}_{G_n}$, we can apply dominated convergence conditionally for

$$\mathbb{E}[\phi(X\mathbf{1}_{G_n}) \mid \mathcal{F}] \rightarrow \mathbb{E}[\phi(X) \mid \mathcal{F}]$$

and we are done. □

LECTURE 9

We continue with properties of conditional expectation.

1. (Tower property) If $\mathcal{G}$ is a sub-sigma-algebra of $\mathcal{F}$ then

$$\mathbb{E}[\mathbb{E}[X | \mathcal{F}] | \mathcal{G}] = \mathbb{E}[X | \mathcal{G}] .$$

Proof. Set $Z = \mathbb{E}[X | \mathcal{G}]$ and $W = \mathbb{E}[X | \mathcal{F}]$. We need to show that

$$\mathbb{E}[W | \mathcal{G}] = Z ,$$

or that for all $A \in \mathcal{G}$,

$$\int_A Z \, d\mathbb{P} = \int_A W \, d\mathbb{P} .$$

But this is the same as

$$\int_A \mathbb{E}[X | \mathcal{G}] \, d\mathbb{P} = \int_A X \, d\mathbb{P} = \int_A \mathbb{E}[X | \mathcal{F}] \, d\mathbb{P} ,$$

which holds since $A \in \mathcal{F}$. □

6.2 Conditional probability

Conditional probability is just a special case of conditional expectation.

Definition 6.5. Let $\mathcal{F}$ be a sub-sigma-algebra of Σ and $A \in \Sigma$ be an event. Define the conditional probability of A given $\mathcal{F}$ as

$$\mathbb{P}(A | \mathcal{F}) = \mathbb{E}[\mathbf{1}_A | \mathcal{F}] .$$

Note that conditional probability is also a function, not a number. The theorem below says that conditional probability acts as a probability measure for almost every ω . We will state a precise version of this later, but at this point we will only prove a functional version.

Theorem 6.6. With probability one,

$$\mathbb{P}(\emptyset | \mathcal{F}) = 0, \mathbb{P}(\Omega | \mathcal{F}) = 1 \text{ and } 0 \leq \mathbb{P}(A | \mathcal{F}) \leq 1 .$$

Furthermore if $A_1, A_2, \dots$ are pairwise disjoint,

$$\mathbb{P}(\cup_n A_n | \mathcal{F}) = \sum_n \mathbb{P}(A_n | \mathcal{F}) .$$

Proof. The first statements are the same as

$$\mathbb{E}[0 | \mathcal{F}] = 0, \mathbb{E}[1 | \mathcal{F}] = 1 \text{ and } 0 \leq \mathbb{E}[\mathbf{1}_A | \mathcal{F}] \leq 1 .$$

These follow from the properties of last section. The second statement uses the fact that if the A_n 's are disjoint, then

$$\mathbb{P}(\cup_n A_n | \mathcal{F}) = \mathbb{E}[\mathbf{1}_{\cup_n A_n} | \mathcal{F}] = \mathbb{E}[\sum_n \mathbf{1}_{A_n} | \mathcal{F}]$$

and by linearity this equals

$$\sum_n \mathbb{E}[\mathbf{1}_{A_n} \mid \mathcal{F}] = \sum_n \mathbb{P}(A_n \mid \mathcal{F}) .$$

□

Here are some more natural properties.

1. If $A \subset B$ then $\mathbb{P}(A \mid \mathcal{F}) \leq \mathbb{P}(B \mid \mathcal{F})$ almost surely.

Proof. This follows from last section, using $\mathbf{1}_A \leq \mathbf{1}_B$ almost surely. □

2. If (A_n) is a nested increasing sequence then $\mathbb{P}(A_n \mid \mathcal{F}) \uparrow \mathbb{P}(\cup_n A_n \mid \mathcal{F})$.

Proof. Use conditional dominated convergence with $|\mathbf{1}_{A_n}| \leq 1$, which is integrable. □

The next theorem says that we can write conditional probability as a true probability measure, for almost every ω . In other words, it allows us to fix almost any ω and claim that $\mathbb{P}(\cdot \mid \mathcal{F})(\omega)$ is a probability measure. It requires some properties of the underlying space Ω , so we will take it to be $\mathbb{R}$.

Theorem 6.7. *Let $\mathbb{P}$ be a Borel probability measure on $\mathbb{R}$ and $\mathcal{F}$ a sub-sigma-algebra of the Borel sets. There exists a function $\mu : \Omega \times \mathcal{B} \rightarrow [0, 1]$, where $\mathcal{B}$ is the Borel subsets of $\mathbb{R}$ such that*

1. $\mu_\omega(\cdot) = \mu(\omega, \cdot) : \mathcal{B} \rightarrow [0, 1]$ is a Borel probability measure on $(\mathbb{R}, \mathcal{F})$ for every $\omega \in \Omega$.
2. For each $A \in \mathcal{B}$,

$$\mu(\cdot, A) = \mathbb{P}(A \mid \mathcal{F}) \text{ almost surely .}$$

Proof. We would like to define a measure $\mu_\omega(A)$ as $\mathbb{P}(A \mid \mathcal{F})(\omega)$ for each A . However there are (usually) uncountably many A 's and these conditional probabilities are only defined almost surely, so it is not clear how to define this simultaneously for all A 's. So we look at a generating collection of A 's, of the form $(-\infty, r]$ for rational r .

For each rational r , define

$$F(r, \omega) = \mathbb{P}((-\infty, r] \mid \mathcal{F}) .$$

Specifically, for fixed r , we define $F(r, \omega)$ to be a *version* of the conditional expectation on the right. That is, it is defined on all of $\Omega = \mathbb{R}$ but agrees with the conditional probability almost surely. We will now work to show that there is a set of probability one on which the $F(r, \omega)$'s define probability measures.

By monotonicity of conditional probability, for each rational $r \leq s$ we can find a set $A_{r,s}$ of probability one such that for $\omega \in A_{r,s}$, we have $F(r, \omega) \leq F(s, \omega)$. Furthermore, for each fixed rational r ,

$$\lim_n F(r + n^{-1}, \omega) = \lim_n \mathbb{P}((-\infty, r + n^{-1}] \mid \mathcal{F}) = \mathbb{P}((-\infty, r] \mid \mathcal{F}) = F(r, \omega)$$

almost surely. Therefore we can find an event B_r of probability one such that for $\omega \in B_r$, we have $F(r + n^{-1}, \omega) \rightarrow F(r, \omega)$. Last, we can find an event C of probability one such that for $\omega \in C$,

$$\lim_{r \rightarrow -\infty} F(r, \omega) = 0 \text{ and } \lim_{r \rightarrow \infty} F(r, \omega) = 1 . \quad (8)$$

Define

$$E = \cap_{r,s} A_{r,s} \bigcap \cap_r B_r \bigcap C .$$

By the above arguments, for $\omega \in E$, which is an event of probability one, the function $r \mapsto F(r, \omega)$ (for rational r) is right-continuous, non-decreasing and satisfies (8). Therefore, just as in the proof of Helly's selection theorem, if we define for $\omega \in E$

$$G(x, \omega) = \inf \{ F(r, \omega) : r > x \} ,$$

then $x \mapsto G(x, \omega)$ is a distribution function of some probability measure μ_ω ; that is,

$$\mu_\omega((-\infty, x]) = G(x, \omega) \text{ for } x \in \mathbb{R} \text{ and } \omega \in E .$$

We must now show that if we define

$$\mu(\omega, A) = \begin{cases} \mu_\omega(A) & \text{if } \omega \in E \\ \delta_0 & \text{if } \omega \notin E \end{cases}$$

then this function satisfies the conditions of the theorem. We will do this next time. □

LECTURE 10

We were in the middle of proving

Theorem 6.8. *Let $\mathbb{P}$ be a Borel probability measure on $\mathbb{R}$ and $\mathcal{F}$ a sub-sigma-algebra of the Borel sets. There exists a function $\mu : \Omega \times \mathcal{B} \rightarrow [0, 1]$, where $\mathcal{B}$ is the Borel subsets of $\mathbb{R}$ such that*

1. $\mu_\omega(\cdot) = \mu(\omega, \cdot) : \mathcal{B} \rightarrow [0, 1]$ is a Borel probability measure on $(\mathbb{R}, \mathcal{F})$ for every $\omega \in \Omega$.
2. For each $A \in \mathcal{B}$,

$$\mu(\cdot, A) = \mathbb{P}(A \mid \mathcal{F}) \text{ almost surely .}$$

Proof. So far we have defined $F(r, \omega)$ as a version:

$$F(r, \omega) = \mathbb{P}((-\infty, r] \mid \mathcal{F}) .$$

On a set E of probability one, we have defined $\mu_\omega(\cdot)$ as a probability measure with distribution function $G(r, \omega)$ that agrees with $F(r, \omega)$ on rationals. We then set

$$\mu(\omega, A) = \begin{cases} \mu_\omega(A) & \text{if } \omega \in E \\ \delta_0 & \text{if } \omega \notin E \end{cases}$$

and must show that this function satisfies the conditions of the theorem.

To do this, we will first prove that

$$\omega \mapsto \mu(\omega, A) \text{ is } \mathcal{F}\text{-measurable for all Borel } A \subset \mathbb{R} . \quad (9)$$

In that vein, first note that the collection $\mathcal{C}$ of Borel A such that $\omega \mapsto \mu(\omega, A)$ is $\mathcal{F}$ -measurable is a λ -system. To check this, first note that $\Omega \in \mathcal{C}$ since

$$\mu(\omega, \Omega) = 1 \text{ for all } \omega$$

and is thus $\mathcal{F}$ -measurable. Furthermore if $A \in \mathcal{C}$ then

$$\mu(\omega, A^c) = \begin{cases} \mu_\omega(A^c) & \text{if } \omega \in E \\ \delta_0(A^c) & \text{if } \omega \in E^c \end{cases} = \begin{cases} 1 - \mu_\omega(A) & \text{if } \omega \in E \\ 1 - \delta_0(A) & \text{if } \omega \in E^c \end{cases} = 1 - \mu(\omega, A)$$

and is $\mathcal{F}$ -measurable. Last, if $A_1, A_2, \dots \in \mathcal{C}$ are pairwise disjoint then

$$\mu(\omega, \cup_n A_n) = \begin{cases} \mu_\omega(\cup_n A_n) & \text{if } \omega \in E \\ \delta_0(\cup_n A_n) & \text{if } \omega \in E^c \end{cases} = \begin{cases} \sum_n \mu_\omega(A_n) & \text{if } \omega \in E \\ \sum_n \delta_0(A_n) & \text{if } \omega \in E^c \end{cases} = \sum_n \mu(\omega, A_n)$$

and again is $\mathcal{F}$ -measurable.

For every rational r , $(-\infty, r] \in \mathcal{C}$. Indeed, for $\omega \in E$,

$$\mu(\omega, (-\infty, r]) = G(\omega, r) = F(\omega, r) ,$$

and this equals $\mathbb{P}((-\infty, r] \mid \mathcal{F})$ almost surely (and this is an $\mathcal{F}$ -measurable function). Since E has probability one, we have shown that $(-\infty, r] \in \mathcal{C}$. The collection of such intervals is a π -system that generates the Borel sigma-algebra, so by the π - λ theorem, $\mathcal{C}$ contains all Borel sets. In other words, (9) holds.

By definition, for $B \in \mathcal{F}$ and $A = (-\infty, r]$ for r rational,

$$\mathbb{P}(A \cap B) = \int_B F(r, \omega) \, d\mathbb{P} = \int_B \mu(\omega, A) \, d\mathbb{P} .$$

For fixed B , both sides of this equation are measures (as a function of A). You can check this for the right side. Since they agree on the π -system of intervals $(-\infty, r]$ for rational r , they must agree on all Borel sets. (Again by π - λ .)

□

6.3 Conditional distributions of random variables

One way to rephrase our result on conditional probability distributions is in terms of random variables. Given a random vector (X, Y) , we can find a family μ_ω of probability measures on $\mathbb{R}$, defined for each $\omega \in \Omega$ such that

$$\mu(\omega, A) = \mathbb{P}(X \in A \mid Y)(\omega) \text{ almost surely .}$$

This family of measures can be called the conditional distribution of X given Y (and can be proved exactly the same way). There is a version for conditional expectations.

Theorem 6.9. *Let (X, Y) be a random vector and $f : \mathbb{R} \rightarrow \mathbb{R}$ be a Borel function such that $\mathbb{E}f(X)$ exists. If μ_ω is a conditional distribution of X given Y , then*

$$\int f(x) \mu(\omega, dx) = \mathbb{E}[f(X) \mid Y] \text{ almost surely.}$$

Proof. If f is the indicator of a Borel set, then this formula holds by the previous result. We can then use linearity to extend to simple functions, then take monotone limits for nonnegative functions. Last, split into positive and negative parts for general functions. $\square$

Billingsley remarks that one can use this theorem to give a short proof of Jensen. Just apply the inequality to the integral above.

LECTURE 11

7 Examples, applications

7.1 Basic probability spaces

We begin with the simplest case. Suppose that $(\Omega, \Sigma, \mathbb{P})$ is a probability space and $\mathcal{F}$ is a sub-sigma-algebra that is generated by a countable partition (A_i) of subsets of Ω . This is the case, for example, when the probability space is finite or countable: $\Omega = \{1, 2, \dots\}$ or when $\mathcal{F}$ is generated by a discrete random variable.

In this case, we expect the new definition of conditional probability to coincide with the old (once we modify it to be referring to functions). So let $A \in \Sigma$ and define

$$f(\omega) = \begin{cases} \frac{\mathbb{P}(A_i \cap A)}{\mathbb{P}(A_i)} & \text{if } \omega \in A_i \text{ and } \mathbb{P}(A_i) > 0 \\ 0 & \text{if } \omega \in A_i \text{ and } \mathbb{P}(A_i) = 0 \end{cases}.$$

Because (A_i) is a partition of Ω , f is defined for all ω . Furthermore, the sets on which f is defined as 0 have probability zero, so the definition there does not really matter. That is, the values there will not affect whether f is a version of $\mathbb{P}(A \mid \mathcal{F})$.

We will now show that $f = \mathbb{P}(A \mid \mathcal{F})$ almost surely. To do this, let $B \in \mathcal{F}$. If $B = A_i$ for some i with $\mathbb{P}(A_i) > 0$, then

$$\int_B f \, d\mathbb{P} = \mathbb{P}(A_i) \frac{\mathbb{P}(A_i \cap A)}{\mathbb{P}(A_i)} = \mathbb{P}(A_i \cap A) = \int_B \mathbf{1}_A \, d\mathbb{P}.$$

Otherwise if $\mathbb{P}(A_i) = 0$ then

$$\int_B f \, d\mathbb{P} = 0 = \int_B \mathbf{1}_A \, d\mathbb{P}.$$

Therefore the above equation holds for all $B \in \{A_i\}$ and since such B 's form a π -system that generate $\mathcal{F}$, it holds for all $B \in \mathcal{F}$. (This is because, as in the last section, the set of B 's for which the above equation holds form a λ -system.)

To finish the proof we must show that f is $\mathcal{F}$ -measurable. But we have already proved in the homework that a function is $\mathcal{F}$ -measurable if and only if it is constant on the partition elements that generate $\mathcal{F}$. This is true for our f , so

$$f = \mathbb{P}(A \mid \mathcal{F}) \text{ almost surely .}$$

Example. Let $X_1, \dots$ be a sequence of simple random variables with values in $\mathbb{R}$. Recall that (X_n) is a Markov chain if for each $n \geq 1$ and each choice of $i_1, \dots, i_n, i_{n+1} \in \mathbb{R}$,

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n, \dots, X_1 = i_1) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n) .$$

(Here we are using the old, “basic” definition of conditional probability. We can relate this Markov property to our current definition of conditional probability. Set $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$. This sigma-algebra is generated by a countable partition. Namely, if R_i is the range of X_i , then we can use the partition

$$\{\{X_1 = i_1, \dots, X_n = i_n\} : i_j \in R_j \text{ for all } j\} .$$

Writing $\mathcal{F}_n^0 = \sigma(X_n)$, we can generate $\mathcal{F}_n^0$ with a different partition

$$\{\{X_n = i_n\} : i_n \in R_n\} .$$

We claim that for Borel $H \subset \mathbb{R}$,

$$\mathbb{P}(X_{n+1} \in H \mid \mathcal{F}_n) = \mathbb{P}(X_{n+1} \in H \mid \mathcal{F}_n^0) \text{ almost surely .}$$

Indeed, any ω in our space is in one of the sets $\{X_1 = i_1, \dots, X_n = i_n\}$, which is a subset of $\{X_n = i_n\}$. For such an ω , if $\mathbb{P}(X_1 = i_1, \dots, X_n = i_n) > 0$,

$$\mathbb{P}(X_{n+1} \in H \mid \mathcal{F}_n) = \mathbb{P}(X_{n+1} \in H \mid X_1 = i_1, \dots, X_n = i_n)$$

and

$$\mathbb{P}(X_{n+1} \in H \mid \mathcal{F}_n^0) = \mathbb{P}(X_{n+1} \in H \mid X_n = i_n) .$$

The Markov property says these are equal.

7.2 The general case

Example. Densities on $\mathbb{R}^2$.

In many basic probability classes, one learns the definition of conditional probability for continuous distributions. We will derive the formula here. Suppose that $f(x, y)$ is a density (relative to Lebesgue measure) for a probability distribution $\mathbb{P}$ on $\mathbb{R}^2$. Let $\mathcal{F}$ be the sigma-algebra generated by the first coordinate x . That is,

$$\mathcal{F} = \sigma(x) = \sigma(\{A \times \mathbb{R} : A \subset \mathbb{R} \text{ Borel}\}) .$$

We claim that for Borel $B \subset \mathbb{R}$,

$$\mathbb{P}(y \in B \mid x) := \mathbb{P}(y \in B \mid \mathcal{F}) = \frac{\int_B f(x, y) \, dy}{\int f(x, y) \, dy} \text{ for } \mathbb{P} \text{ a.e. } x .$$

In other words, to obtain the conditional probability distribution, we just plug in the value of x and normalize.

To prove this, let $A \subset \mathbb{R}$ be Borel and define ϕ by

$$\phi(x, y) = \begin{cases} \frac{\int_B f(x, y) \, dy}{\int f(x, y) \, dy} & \text{if } \int f(x, y) \, dy > 0 \\ 0 & \text{if } \int f(x, y) \, dy = 0 \end{cases} .$$

Note that this function only depends on x , so it is measurable relative to $\mathcal{F}$. We must then show

$$\mathbb{E} \mathbf{1}_{A \times \mathbb{R}} \phi(x, y) = \mathbb{E} \mathbf{1}_{A \times \mathbb{R}} \mathbf{1}_{\mathbb{R} \times B} .$$

The left side is, by Fubini,

$$\int_A \int \phi(x, y) f(x, y) \, dy \, dx = \int_A \left[\int_B f(x, y) \, dy \right] \, dx .$$

And this equals the right side.

LECTURE 12

More examples of conditional probability:

Example. Here is a generalization of the last example. Let X, Y be independent random vectors of dimensions n_1 and n_2 respectively, with $n_1 + n_2 = n$. Then, by Fubini, the distribution of (X, Y) is a product measure $\mu = \mu_1 \times \mu_2$, where μ_1 is a probability measure on $\mathbb{R}^{n_1}$ and μ_2 is a probability measure on $\mathbb{R}^{n_2}$, and we proved last semester that

$$\mu(A) = \int \mu_1 \{x \in \mathbb{R}^{n_1} : (x, Y(\omega)) \in A\} \, d\mathbb{P}, \quad A \subset \mathbb{R}^n \text{ Borel} .$$

We will show that this implies that

$$\mathbb{P}((X, Y) \in A \mid Y)(\omega) = \mu_1 \{x \in \mathbb{R}^{n_1} : (x, Y(\omega)) \in A\} .$$

In other words, we fix the value of Y to be $y = Y(\omega)$ and take the probability of the set of ω' such that $(X(\omega'), y) \in A$.

To do this, we must show that for $B \subset \mathbb{R}^{m_2}$ Borel,

$$\int_{\{Y \in B\}} \mathbf{1}_{\{(X, Y) \in A\}} \, d\mathbb{P} = \int_{\{Y \in B\}} \mu_1 \{x \in \mathbb{R}^{n_1} : (x, Y(\omega)) \in A\} \, d\mathbb{P} .$$

The left side equals

$$\mu(A \cap (\mathbb{R}^{m_1} \times B)) .$$

On the other hand, the right side equals

$$\int \mu_1\{x \in \mathbb{R}^{n_1} : (x, Y(\omega)) \in A\} \mathbf{1}_{\{Y(\omega) \in B\}} d\mathbb{P} .$$

Now for all ω ,

$$\mu_1\{x \in \mathbb{R}^{n_1} : (x, Y(\omega)) \in A\} \mathbf{1}_{\{Y(\omega) \in B\}} = \mu_1\{x \in \mathbb{R}^{n_1} : (x, Y(\omega)) \in A \cap (\mathbb{R}^{n_1} \times B)\} .$$

So we get

$$\int \mu_1\{x \in \mathbb{R}^{n_1} : (x, Y(\omega)) \in A \cap (\mathbb{R}^{n_1} \times B)\} d\mathbb{P} = \mu(A \cap (\mathbb{R}^{n_1} \times B)) ,$$

which is the left side. In other words, if we “condition on $Y = y$ ” for independent random vectors X, Y , then we can just place the value y in for Y .

Example. (Markov process) A collection of random variables $(X_t : t \geq 0)$ is called a (continuous time) Markov process if for all $t \leq t'$ and Borel $H \subset \mathbb{R}$,

$$\mathbb{P}(X_{t'} \in H \mid X_s, 0 \leq s \leq t) = \mathbb{P}(X_{t'} \in H \mid X_t) \text{ almost surely} . \quad (10)$$

The sigma-algebra on the left is the one generated by all X_s ’s up to time t ; that is, it is

$$\sigma(X_s : 0 \leq s \leq t) .$$

One can show that this Markov property is equivalent to the following: for every $0 \leq t_1 < t_2 < \dots < t_k = t \leq t'$, one has

$$\mathbb{P}(X_{t'} \in H \mid X_{t_1}, \dots, X_{t_k}) = \mathbb{P}(X_{t'} \in H \mid X_{t_k}) \text{ almost surely} . \quad (11)$$

To prove that (10) is equivalent to (11), first assume (10). Set $A = \{X_{t'} \in H\}$, $\mathcal{F}_1 = \sigma(X_s : 0 \leq s \leq t)$ and for fixed $0 \leq t_1 < t_2 < \dots < t_k = t \leq t'$, set $\mathcal{F}_2 = \sigma(X_{t_1}, \dots, X_{t_k})$. By the Tower property, since $\mathcal{F}_2 \subset \mathcal{F}_1$,

$$\mathbb{P}(A \mid \mathcal{F}_2) = \mathbb{E}(\mathbb{P}(A \mid \mathcal{F}_1) \mid \mathcal{F}_2) = \mathbb{E}(\mathbb{P}(A \mid X_t) \mid \mathcal{F}_2) = \mathbb{P}(A \mid X_t) .$$

Conversely, assume (11). Note that the union

$$\bigcup_{0 \leq t_1 < \dots < t_k = t} \sigma(X_{t_1}, \dots, X_{t_k})$$

is a π -system that generates $\sigma(X_s : 0 \leq s \leq t)$. Furthermore by (11), for each A in this union,

$$\int_A \mathbf{1}_{\{X_{t'} \in H\}} d\mathbb{P} = \int_A \mathbb{P}(X_{t'} \in H \mid X_t) d\mathbb{P} .$$

Because the set of A which satisfy this identity forms a λ -system, it must then hold for all $A \in \sigma(X_s : 0 \leq s \leq t)$ and we have shown (10).

As an example, we can show that a Poisson process is a continuous time Markov process. Recall that N_t is defined as follows: let $X_1, X_2, \dots$ be an i.i.d. sequence of exponential random variables with parameter $\alpha > 0$. Then we set $X_0 = 0$ and

$$N_t = \min\{n \geq 0 : X_0 + \dots + X_n \geq t\} .$$

We showed last semester that

1. for $0 \leq t_1 < t_2 < \dots < t_k$, the increments $N_{t_2} - N_{t_1}, \dots, N_{t_k} - N_{t_{k-1}}$ are independent and
2. for $0 \leq s \leq t$, $N_t - N_s$ has Poisson distribution with parameter $\alpha(t - s)$.

Using these results, we can show the Markov property by verifying (11). Let $0 \leq t_1 < \dots < t_k = t \leq t'$ and let $H \subset \mathbb{R}$ be Borel. Then

$$\mathbb{P}(N_{t'} \in H \mid N_{t_1}, \dots, N_{t_k}) = \mathbb{P}((N_{t_1}, \dots, N_{t_k}, N_{t'} - N_t) \in A \mid N_{t_1}, \dots, N_{t_k}) ,$$

where $A \subset \mathbb{R}^{k+1}$ is the set of points $(x_1, \dots, x_{k+1})$ such that $x_{k+1} + x_k \in H$. Because $N_{t'} - N_t$ is independent of the random vector $(N_{t_1}, \dots, N_{t_k})$, we can use the result of a previous example to write this as a function of ω as

$$\mathbb{P}(\omega' : (N_{t_1}(\omega), \dots, N_{t_k}(\omega), N_{t'}(\omega') - N_t(\omega')) \in A)$$

and this equals

$$\mathbb{P}(\omega' : (N_{t_k}(\omega), N_{t'}(\omega') - N_t(\omega')) \in \hat{A}) ,$$

where $\hat{A} \subset \mathbb{R}^2$ is the set of (x_1, x_2) such that $x_1 + x_2 \in H$. Again, by independence, this equals

$$\mathbb{P}((N_t, N_{t'} - N_t) \in A \mid N_t) = \mathbb{P}(N_{t'} \in H \mid N_t) .$$

LECTURE 13

8 Martingales

Martingales can be viewed in many different ways and the first view we will take is that they are an extension of independent sums. To see their defining property, let $\Delta_1, \Delta_2, \dots$ be independent and with mean zero. We will think as in the gambling example that Δ_n is the amount of money we win at time n (that is, at the n -th round of our gambling game) when we bet 1 dollar per round. Then our cumulative wealth at time n is

$$X_n = \Delta_1 + \dots + \Delta_n \text{ with } X_0 = 0 .$$

Now we will use the theory of conditional probability. Let $\mathcal{F}_0 = \{\emptyset, \Omega\}$ be the trivial sigma algebra (the information we have at time 0) and

$$\mathcal{F}_n = \sigma(X_1, \dots, X_n) \text{ for } n \geq 1$$

be the information we have at time n . We can then compute the conditional expectation

$$\begin{aligned}\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] &= \mathbb{E}[X_n + \Delta_{n+1} \mid \mathcal{F}_n] \\ &= \mathbb{E}[X_n \mid \mathcal{F}_n] + \mathbb{E}[\Delta_{n+1} \mid \mathcal{F}_n] .\end{aligned}$$

Now X_n is $\mathcal{F}_n$ -measurable and Δ_{n+1} is independent of $\mathcal{F}_n$. So this reduces to

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = X_n + \mathbb{E}\Delta_{n+1} = X_n .$$

The way to interpret this is that the game is *fair*; that is, given our information up to time n , our best guess for our wealth at time $n + 1$ is just what we currently have: X_n .

A Martingale is just a generalization of a fair game.

Definition 8.1. A sequence $(X_n, \mathcal{F}_n)$ of random variables and sigma-algebras is called a *martingale* (or (X_n) is a martingale relative to $(\mathcal{F}_n)$) if the following conditions are met.

1. $\mathbb{E}|X_n| < \infty$ for all n .
2. X_n is $\mathcal{F}_n$ -measurable for all n .
3. $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = X_n$ almost surely for all n .
4. $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ for all n .

We say (X_n) is a martingale if it is a martingale relative to some sequence $(\mathcal{F}_n)$.

Here are some remarks on the definition.

- The condition $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ makes $(\mathcal{F}_n)$ a *filtration*. Because X_n is $\mathcal{F}_n$ -measurable for all n , we say that (X_n) is *adapted to the filtration* $(\mathcal{F}_n)$.
- If $(X_n, \mathcal{F}_n)$ is a martingale then also $(X_n, \mathcal{G}_n)$ is a martingale, where

$$\mathcal{G}_n = \sigma(X_1, \dots, X_n) .$$

The only condition we need to check is that

$$\mathbb{E}[X_{n+1} \mid X_1, \dots, X_n] = X_n \text{ almost surely .}$$

This holds by the tower property. Indeed, since each of $X_1, \dots, X_n$ are $\mathcal{F}_n$ -measurable, it follows that $\sigma(X_1, \dots, X_n) \subset \mathcal{F}_n$. Therefore

$$\begin{aligned}\mathbb{E}[X_{n+1} \mid X_1, \dots, X_n] &= \mathbb{E}[\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \mid X_1, \dots, X_n] \\ &= \mathbb{E}[X_n \mid X_1, \dots, X_n] \\ &= X_n .\end{aligned}$$

- By the Tower property again, for $n, k \geq 1$,

$$\mathbb{E}[X_{n+k} \mid \mathcal{F}_n] = \mathbb{E}[\mathbb{E}[X_{n+k} \mid \mathcal{F}_{n+k-1}] \mid \mathcal{F}_n] = \mathbb{E}[X_{n+k-1} \mid \mathcal{F}_n] .$$

Iterating this argument,

$$\mathbb{E}[X_{n+k} \mid \mathcal{F}_n] = \mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = X_n .$$

- For each n ,

$$\mathbb{E}X_n = \mathbb{E}[\mathbb{E}[X_{n+1} \mid \mathcal{F}_n]] = \mathbb{E}X_{n+1} .$$

Therefore

$$\mathbb{E}X_n = \mathbb{E}X_0 \text{ for all } n .$$

- A *martingale difference sequence* is defined as

$$\Delta_n = X_n - X_{n-1} \text{ for } n \geq 1 ,$$

where (X_n) is a martingale. If X_n is the generalization of an i.i.d. sum, then Δ_n is the generalization of the i.i.d. summand. So, to be expected,

$$\mathbb{E}\Delta_n = \mathbb{E}X_n - \mathbb{E}X_{n-1} = 0 .$$

Furthermore the martingale property for X_n is equivalent to

$$\mathbb{E}[\Delta_{n+1} \mid \mathcal{F}_n] = X_n - X_n = 0 \text{ almost surely .}$$

Last, if (X_n) is a martingale such that $\mathbb{E}X_n^2 < \infty$ for all n then the Δ_n 's are uncorrelated (as in an i.i.d. sum). To see this we use the tower property. Take $j < k$ and compute

$$\begin{aligned} \mathbb{E}\Delta_j\Delta_k &= \mathbb{E}[\mathbb{E}[\Delta_j\Delta_k \mid \mathcal{F}_{k-1}]] \\ &= \mathbb{E}[\Delta_j\mathbb{E}[\Delta_k \mid \mathcal{F}_{k-1}]] \\ &= 0 . \end{aligned}$$

Here we have used that Δ_j is $\mathcal{F}_{k-1}$ -measurable and that $\mathbb{E}[\Delta_k \mid \mathcal{F}_{k-1}] = 0$.

Examples.

1. Let X be integrable and let $(\mathcal{F}_n)$ be a filtration. Then defining $X_n = \mathbb{E}[X \mid \mathcal{F}_n]$, we can show that $(X_n, \mathcal{F}_n)$ is a martingale. Indeed,

$$\mathbb{E}|X_n| = \mathbb{E}|\mathbb{E}[X \mid \mathcal{F}_n]| \leq \mathbb{E}|X| < \infty ,$$

and X_n is $\mathcal{F}_n$ measurable. Last, by the tower property,

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{F}_{n+1}] \mid \mathcal{F}_n] = \mathbb{E}[X \mid \mathcal{F}_n] = X_n .$$

So we can always build a martingale by conditioning one variable on a filtration. This is called a *Doob martingale*. It will turn out that in wide generality, we can express a martingale X_n as $\mathbb{E}[X \mid \mathcal{F}_n]$ for some random variable X .

2. Let f be a Lebesgue integrable function on $[0, 1]$. Then define f_n by

$$f_n(x) = 2^n \int_{\frac{j-1}{2^n}}^{\frac{j}{2^n}} f(t) dt \text{ for } t \in \left[\frac{j-1}{2^n}, \frac{j}{2^n} \right) .$$

Then on the probability space $[0, 1]$ with Lebesgue measure, f_n is just $\mathbb{E}[f \mid \mathcal{F}_n]$, where $\mathcal{F}_n$ is generated by the dyadic intervals. Such a sequence is called a *dyadic martingale*. There is an interesting interplay between martingale theory and harmonic analysis brought about by analyzing dyadic martingales.

3. (Branching process) Imagine that at time 1, we have one organism. At time 2, that organism dies and gives birth to a random number Z_2 of organisms. At time 3, each one of these dies and gives birth to a random number of organisms, the total number being Z_3 . We can model this as follows. Let $(N_{n,k} : n, k \geq 1)$ be an i.i.d. array of integer valued random variables. Set $Z_1 = 1$ and inductively,

$$Z_n = N_{n-1,1} + \cdots + N_{n-1,Z_{n-1}} .$$

Here $N_{n-1,k}$ is the number of children that the k -th organism (who was born at time $n-1$) has.

Suppose that $\mathbb{E}N_{n,k} = m$. Then we claim that $Y_n := Z_n/m^n$ is a martingale (relative to its own filtration). To see why, calculate

$$\begin{aligned} \mathbb{E}[Z_n \mid Z_1, \dots, Z_{n-1}] &= \mathbb{E}[N_{n-1,1} + \cdots + N_{n-1,Z_{n-1}} \mid Z_1, \dots, Z_{n-1}] \\ &= \sum_{k=0}^{\infty} \mathbb{E}[(N_{n-1,1} + \cdots + N_{n-1,Z_{n-1}}) \mathbf{1}_{\{Z_{n-1}=k\}} \mid Z_1, \dots, Z_{n-1}] \\ &= \sum_{k=0}^{\infty} \mathbf{1}_{\{Z_{n-1}=k\}} \mathbb{E}[(N_{n-1,1} + \cdots + N_{n-1,k}) \mid Z_1, \dots, Z_{n-1}] \\ &= \sum_{k=0}^{\infty} \mathbf{1}_{\{Z_{n-1}=k\}} km = Z_{n-1}m . \end{aligned}$$

Therefore

$$\mathbb{E}[Y_n \mid Y_1, \dots, Y_{n-1}] = m^{-n} \mathbb{E}[Z_n \mid Z_1, \dots, Z_{n-1}] = Y_{n-1} .$$

LECTURE 14

8.1 Submartingales and supermartingales

The definition of submartingale is similar:

Definition 8.2. A sequence $(X_n, \mathcal{F}_n)$ is a submartingale if

1. $\mathbb{E}|X_n| < \infty$ for all n .

2. X_n is $\mathcal{F}_n$ -measurable for all n .
3. $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \geq X_n$ almost surely for all n .
4. $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ for all n .

A submartingale represents a gambling game which is at least fair (that is, it may be in your favor, but not against you). Here are some corresponding remarks.

- $\mathbb{E}X_1 \leq \mathbb{E}X_2 \leq \dots$
- $\mathbb{E}[X_{n+k} \mid \mathcal{F}_n] \geq X_n$.
- If X_n is a martingale and $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is convex with $\mathbb{E}\phi(X_n) < \infty$ for all n , then $\phi(X_n)$ is a submartingale.

Proof. By Jensen,

$$\mathbb{E}[\phi(X_{n+1}) \mid \mathcal{F}_n] \geq \phi(\mathbb{E}[X_{n+1} \mid \mathcal{F}_n]) = \phi(X_n) .$$

□

A supermartingale has the same definition, but with item 4 replaced by

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \leq X_n \text{ almost surely for all } n .$$

There are corresponding remarks (as above, but with signs flipped).

8.2 Gambling systems revisited

As in the motivation to martingales above, suppose we play a gambling game where we begin with some fortune X_0 and at time n our fortune is X_n . Then in the case of a fair game it is natural to assert that X_n is a martingale. So $\Delta_n = X_n - X_{n-1}$ for $n \geq 1$ is the amount won at the n -th game (assuming that 1 dollar is bet). If instead of betting one dollar, one bets W_n dollars, then the total wealth at time n is

$$Y_n = X_0 + W_1\Delta_1 + \dots + W_n\Delta_n .$$

We assume as before that the wager W_n made at time n is bounded and depends only on information up to time $n - 1$; that is W_n is $\mathcal{F}_n$ -measurable (where $\mathcal{F}_n = \sigma(X_0, \dots, X_n)$). Then

Y_n is still a martingale .

To see why, just compute

$$\begin{aligned} \mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] &= \mathbb{E}[Y_n \mid \mathcal{F}_n] + \mathbb{E}[W_{n+1}\Delta_{n+1} \mid \mathcal{F}_n] \\ &= Y_n + W_{n+1}\mathbb{E}[\Delta_{n+1} \mid \mathcal{F}_n] = Y_n . \end{aligned}$$

This means that if the game is fair and we make any wagers we want, it is still fair. That is, we cannot get any advantage from making strategic bets.

We can also introduce a stopping time, as in last semester. It is a function

$$\tau : \Omega \rightarrow \mathbb{N} \cup \{\infty\} \text{ satisfying } \{\tau = n\} \in \mathcal{F}_n \text{ for all } n .$$

(One can also define as $\{\tau \leq n\} \in \mathcal{F}_n$ for all n ; this is equivalent.) This is a function telling us when to stop betting. That is, we define

$$X_n^* = \begin{cases} X_n & \text{if } n \leq \tau \\ X_\tau & \text{if } n \geq \tau \end{cases} .$$

This will still be a fair game, meaning that X_n^* is still a martingale, as we will show. To verify the definition, first

$$\begin{aligned} \mathbb{E}|X_n^*| &= \sum_{k=0}^{\infty} \mathbb{E}|X_n^*| \mathbf{1}_{\{\tau=k\}} = \sum_{k=0}^{n-1} \mathbb{E}|X_k| \mathbf{1}_{\{\tau=k\}} + \mathbb{E}|X_n| \mathbf{1}_{\{\tau \geq n\}} \\ &\leq \sum_{k=0}^{n-1} \mathbb{E}|X_k| + \mathbb{E}|X_n| < \infty . \end{aligned}$$

Next we show that X_n^* is $\mathcal{F}_n$ -measurable. For $H \subset \mathbb{R}$ Borel,

$$\{X_n^* \in H\} = \left[\bigcup_{k=0}^{n-1} \{X_k \in H, \tau = k\} \right] \cup \{X_n \in H, \tau \geq n\} .$$

Since X_k is $\mathcal{F}_k$ -measurable, the definition of stopping time gives

$$\{X_k \in H, \tau = k\} \in \mathcal{F}_k \subset \mathcal{F}_n ,$$

so the first union is in $\mathcal{F}_n$. For the second,

$$\{\tau \geq n\} = \left[\bigcup_{k=0}^{n-1} \{\tau = k\} \right]^c \in \mathcal{F}_{n-1} \subset \mathcal{F}_n ,$$

so since X_n is $\mathcal{F}_n$ -measurable, the second term of the union is also in $\mathcal{F}_n$ and we conclude that X_n^* is $\mathcal{F}_n$ -measurable.

Last,

$$\begin{aligned} \mathbb{E}[X_{n+1}^* \mid \mathcal{F}_n] &= \sum_{k=0}^n \mathbb{E}[X_k \mathbf{1}_{\{\tau=k\}} \mid \mathcal{F}_n] + \mathbb{E}[X_{n+1} \mathbf{1}_{\{\tau \geq n+1\}} \mid \mathcal{F}_n] \\ &= \sum_{k=0}^n X_k \mathbf{1}_{\{\tau=k\}} + \mathbf{1}_{\{\tau \geq n+1\}} \mathbb{E}[X_{n+1} \mid \mathcal{F}_n] . \end{aligned}$$

Here we have used that the terms $\mathbf{1}_{\{\tau=k\}}$, X_k and $\mathbf{1}_{\{\tau \geq n+1\}}$ are $\mathcal{F}_n$ -measurable. By the martingale property, this equals

$$\sum_{k=0}^n X_k \mathbf{1}_{\{\tau=k\}} + X_n \mathbf{1}_{\{\tau \geq n+1\}} = X_n^* .$$

We conclude yet again that we cannot gain an advantage by stopping when we want.

8.3 Stopping times and the stopped sigma-algebra

The stopped sigma-algebra is defined, for a stopping time τ , by

$$\mathcal{F}_\tau = \{A \in \Sigma : A \cap \{\tau = k\} \in \mathcal{F}_k \text{ for all } k\} .$$

This definition is quite weird, but it is supposed to be the information that we have at the time we stop playing. To understand the definition, we can partition the space Ω given by the events $\{\tau = k\}$ (imagine here that $\tau < \infty$). On the portion of the space covered by $\{\tau = k\}$, $\mathcal{F}_\tau$ should represent the information given in $\mathcal{F}_k$. So an event should be knowable at time τ if on each set $\{\tau = k\}$, the piece of A in this set is in $\mathcal{F}_k$.

- $\mathcal{F}_\tau$ can be equivalently defined as

$$\mathcal{F}_\tau = \{A \in \Sigma : A \cap \{\tau \leq k\} \in \mathcal{F}_k \text{ for all } k\} .$$

Proof. If A satisfies $A \cap \{\tau \leq k\} \in \mathcal{F}_k$ for all k then

$$A \cap \{\tau = k\} = A \cap (\{\tau \leq k\} \setminus \{\tau \leq k-1\}) = (A \cap \{\tau \leq k\}) \setminus (A \cap \{\tau \leq k-1\}) ,$$

which is in $\mathcal{F}_k$, since the first is in $\mathcal{F}_k$ and the second in $\mathcal{F}_{k-1} \subset \mathcal{F}_k$.

Conversely, if A satisfies $A \cap \{\tau = k\} \in \mathcal{F}_k$ for all k , then

$$A \cap \{\tau \leq k\} = \bigcup_{j=1}^k (A \cap \{\tau = j\}) \in \mathcal{F}_k .$$

□

LECTURE 15

Recall that if τ is a stopping time, then the stopped sigma-algebra is defined by

$$\mathcal{F}_\tau = \{A \in \Sigma : A \cap \{\tau = k\} \in \mathcal{F}_k \text{ for all } k\} .$$

- $\mathcal{F}_\tau$ is a sigma-algebra.

Proof. First $\Omega \in \mathcal{F}_\tau$ since $\{\tau = k\} \in \mathcal{F}_k$ for all k . Next if $A \in \mathcal{F}_\tau$ then for each k , $A^c \cap \{\tau = k\} = \{\tau = k\} \setminus (A \cap \{\tau = k\})$ and this is in $\mathcal{F}_k$. Last if $A_1, A_2, \dots \in \mathcal{F}_\tau$ then for each k ,

$$(\bigcup_n A_n) \cap \{\tau = k\} = \bigcup_n (A_n \cap \{\tau = k\}) \in \mathcal{F}_k .$$

□

- τ is $\mathcal{F}_\tau$ -measurable.

Proof. For any n ,

$$\{\tau = n\} \cap \{\tau = k\} = \begin{cases} \emptyset & \text{if } n \neq k \\ \{\tau = k\} & \text{if } n = k \end{cases} .$$

So in all cases, this is in $\mathcal{F}_k$. Therefore $\{\tau = n\} \in \mathcal{F}_\tau$. Since τ takes values in $\mathbb{N} \cup \{\infty\}$, this suffices to show that τ is $\mathcal{F}_\tau$ -measurable. $\square$

- If τ is identically n then $\mathcal{F}_\tau = \mathcal{F}_n$.

Proof. Suppose that $\tau = n$. Then $\{\tau = k\} = \emptyset$ when $k \neq n$ so in these cases all events A satisfy $A \cap \{\tau = k\} \in \mathcal{F}_k$. Therefore

$$\mathcal{F}_\tau = \{A : A \cap \{\tau = n\} \in \mathcal{F}_n\} .$$

But $\{\tau = n\} = \Omega$ so this is $\mathcal{F}_n$. $\square$

- If $\tau_1 \leq \tau_2$ pointwise then $\mathcal{F}_{\tau_1} \subset \mathcal{F}_{\tau_2}$.

Proof. Assume $\tau_1 \leq \tau_2$ and let $A \in \mathcal{F}_{\tau_1}$. Then for any k ,

$$A \cap \{\tau_2 = k\} = \cup_{j=0}^k [A \cap \{\tau_1 = j\} \cap \{\tau_2 = k\}] .$$

Since $A \cap \{\tau_1 = j\} \in \mathcal{F}_j \subset \mathcal{F}_k$ and $\{\tau_2 = k\} \in \mathcal{F}_k$, this is in $\mathcal{F}_k$ and we are done. $\square$

There are theorems that state that for martingales (and also sometimes for Markov processes), stopping times behave like times. One example is the optional stopping theorem. In the martingale case, it says roughly that there is nothing to gain by stopping at a stopping time.

Theorem 8.3 (Optional stopping theorem). *Let $(X_1, \mathcal{F}_1), \dots, (X_n, \mathcal{F}_n)$ be a submartingale and let $1 \leq \tau_1 \leq \tau_2 \leq n$ be stopping times. Then*

$$\mathbb{E}[X_{\tau_2} \mid \mathcal{F}_{\tau_1}] \geq X_{\tau_1} .$$

If submartingale is changed to martingale then $\geq$ is changed to $=$.

Another way to state the theorem is that the submartingale (or martingale) property holds for bounded stopping times.

Proof. First

$$\mathbb{E}|X_{\tau_2}| \leq \sum_{k=1}^n \mathbb{E}|X_k| < \infty ,$$

so it is integrable. Note that $\mathbb{E}[X_{\tau_2} \mid \mathcal{F}_{\tau_1}]$ and X_{τ_1} are $\mathcal{F}_{\tau_1}$ -measurable, so to show the theorem, we can just show that

$$\int_A (\mathbb{E}[X_{\tau_2} \mid \mathcal{F}_{\tau_1}] - X_{\tau_1}) \, d\mathbb{P} \geq 0 \text{ for } A \in \mathcal{F}_{\tau_1} . \quad (12)$$

This is equivalent to

$$\int_A (X_{\tau_2} - X_{\tau_1}) \, d\mathbb{P} \geq 0 .$$

We then write, using $\tau_1 \leq \tau_2$,

$$X_{\tau_2} - X_{\tau_1} = \sum_{k=\tau_1+1}^{\tau_2} \Delta_k = \sum_k \mathbf{1}_{\{\tau_1 < k \leq \tau_2\}} \Delta_k ,$$

so the above integral is

$$\sum_k \int \mathbf{1}_{A \cap \{\tau_1 < k \leq \tau_2\}} \Delta_k \, d\mathbb{P} .$$

Note that since τ_1 and τ_2 are stopping times,

$$A \cap \{\tau_1 < k \leq \tau_2\} = [A \cap \{\tau_1 \leq k-1\}] \cap \{\tau_2 \geq k\} \in \mathcal{F}_{k-1} .$$

So we can replace by

$$\sum_k \int \mathbf{1}_{A \cap \{\tau_1 < k \leq \tau_2\}} \mathbb{E}[\Delta_k \mid \mathcal{F}_{k-1}] \, d\mathbb{P} ,$$

which is nonnegative by the submartingale property.

In the case that X_n is a martingale, the inequality in (12) is equality and this leads to $\mathbb{E}[X_{\tau_2} \mid \mathcal{F}_{\tau_1}] = X_{\tau_1}$. $\square$

9 Applications of stopping times

As applications, we give two important inequalities. The first is an extension of Komogorov's maximal inequality for i.i.d. sums to the martingale setting. Note below that if (X_n) is actually a martingale, $(|X_n|)$ is a submartingale, so the inequality below holds for $\max_{1 \leq i \leq n} |X_i|$.

Theorem 9.1 (Maximal inequality). *If (X_n) is a submartingale then for $\alpha > 0$,*

$$\mathbb{P} \left[\max_{1 \leq i \leq n} X_i \geq \alpha \right] \leq \frac{1}{\alpha} \mathbb{E}[X_n] .$$

Proof. Set $M_k = \max_{1 \leq i \leq k} X_i$. Define σ by

$$\sigma = \min\{k = 1, \dots, n : X_k \geq \alpha\}$$

and $\tau = \min\{\sigma, n\}$. Then τ is a stopping time (and so is σ):

$$\{\tau = k\} = \begin{cases} \{X_1 < \alpha, \dots, X_{k-1} < \alpha, X_k \geq \alpha\} & \text{if } k < n \\ \{M_{n-1} < \alpha\} & \text{if } k = n \end{cases} \in \mathcal{F}_k ,$$

where $(\mathcal{F}_k)$ is the filtration generated by the X_k 's.

Note that $\{M_n \geq \alpha\} \in \mathcal{F}_\tau$. This is because for $k \leq n$,

$$\{M_n \geq \alpha\} \cap \{\tau \leq k\} = \{M_k \leq \alpha\} ,$$

which is in $\mathcal{F}_k$. For $k > n$ the event $\{\tau \leq k\}$ is just the event Ω , so $\{M_n \geq \alpha\} \cap \{\tau \leq k\} \in \mathcal{F}_k$ in that case too. So we can use the optional stopping theorem for $\mathbb{E}[X_n \mid \mathcal{F}_\tau] \geq X_\tau$ and deduce

$$\begin{aligned} \alpha \mathbb{P}(M_n \geq \alpha) &\leq \int_{\{M_n \geq \alpha\}} X_\tau \, d\mathbb{P} \\ &\leq \int_{\{M_n \geq \alpha\}} \mathbb{E}[X_n \mid \mathcal{F}_\tau] \, d\mathbb{P} = \int_{\{M_n \geq \alpha\}} X_n \, d\mathbb{P} \leq \mathbb{E}[X_n] . \end{aligned}$$

The first inequality holds because $X_\tau \geq \alpha$ when $M_n \geq \alpha$. □

LECTURE 16

The next inequality is useful for proving martingale convergence theorems. For it, we need to define a sequence of *up-crossing times*.

Definition 9.2. *If $X_1, X_2, \dots$ is a sequence of random variables, we say (X_n) has an up-crossing of the interval $[a, b]$ between times $m < n$ if*

$$X_m \leq a \text{ but } X_n \geq b .$$

We would like to count the number of up-crossings of the interval $[a, b]$ for a martingale sequence (X_n) between times $1, \dots, n$. To do this, we introduce stopping times.

$$\tau_1 = \begin{cases} \min\{i \in [1, n] : X_i \leq a\} & \text{if this set is nonempty} \\ n & \text{otherwise} \end{cases} .$$

This is the first time the sequence drops below a (if it does). Then we set

$$\tau_2 = \begin{cases} \min\{i \in [1, n] : i > \tau_1, X_i \geq b\} & \text{if this set is nonempty} \\ n & \text{otherwise} \end{cases} .$$

The time interval $[\tau_1, \tau_2]$ corresponds to an up-crossing of $[a, b]$ exactly when $X_{\tau_1} \leq a$ and $X_{\tau_2} \geq b$.

We continue and define inductively:

$$\text{for } k \geq 1, \tau_{2k} = \begin{cases} \min\{i \in [1, n] : i > \tau_{2k-1}, X_i \geq b\} & \text{if this set is nonempty} \\ n & \text{otherwise} \end{cases}$$

and

$$\text{for } k \geq 1, \tau_{2k+1} = \begin{cases} \min\{i \in [1, n] : i > \tau_{2k}, X_i \leq a\} & \text{if this set is nonempty} \\ n & \text{otherwise} \end{cases} .$$

You can verify that all the τ_i 's are stopping times. We then define the number of up-crossings of $[a, b]$ with $a < b$ until time n as

$$U_{a,b}^{(n)} = \max\{i \geq 1 : X_{\tau_{2i-1}} \leq a < b \leq X_{\tau_{2i}}\} .$$

If the sequence (X_n) has many up-crossings of an interval, this means that the sequence fluctuates. Therefore to prove that (X_n) converges, we would like to bound the number of up-crossings of intervals.

Theorem 9.3 (Up-crossing inequality). *If (X_n) is a submartingale then for $a < b$,*

$$\mathbb{E}U_{a,b}^{(n)} \leq \frac{\mathbb{E}|X_n| + |a|}{b - a} .$$

Proof. First we will prove a special case. Suppose that $a = 0$ and $X_n \geq 0$ for all n . Then, since $\tau_i < \tau_{i+1}$ whenever $\tau_i < n$, we must have $\tau_n = n$. Thus we can write

$$X_n = X_{\tau_n} \geq X_{\tau_n} - X_{\tau_1} = \sum_{k=2}^n [X_{\tau_k} - X_{\tau_{k-1}}] .$$

Write Σ_e for the sum over the even k 's and Σ_o for the sum over odd k 's. Note that all the up-crossings of $[a, b] = [0, b]$ are recorded in Σ_e and each time there is an up-crossing, the difference $X_{\tau_k} - X_{\tau_{k-1}}$ is at least $b - a$. Furthermore for intervals $[\tau_{k-1}, \tau_k]$ (with k even) that do not correspond to an up-crossing we have either $\tau_{k-1} = \tau_k = n$ or $\tau_{k-1} < n = \tau_k$ and in both of these scenarios $X_{\tau_k} - X_{\tau_{k-1}} \geq 0$ (by non negativity of the X_i 's). Therefore,

$$\Sigma_e \geq (b - a)U_{a,b}^{(n)} .$$

Again by non-negativity of the X_n 's,

$$\mathbb{E}|X_n| = \mathbb{E}\Sigma_e + \mathbb{E}\Sigma_o \geq (b - a)\mathbb{E}U_{a,b}^{(n)} + \mathbb{E}\Sigma_o .$$

On the other hand, for optional stopping and the fact that (X_n) is a submartingale, when k is odd,

$$\mathbb{E}(X_{\tau_k} - X_{\tau_{k-1}}) \geq \mathbb{E}X_{\tau_{k-1}} - X_{\tau_{k-1}} = 0 .$$

Therefore $\mathbb{E}\Sigma_o \geq 0$ and we obtain

$$\mathbb{E}U_{a,b}^{(n)} \leq \frac{\mathbb{E}|X_n|}{b-a} = \frac{\mathbb{E}|X_n| + |a|}{b-a} .$$

This proves the theorem in our special case.

For the general case, let (X_n) be a submartingale and suppose that a is not necessarily 0. Then define Y_k by

$$Y_k = \max\{0, X_k - a\} = \phi(X_k) ,$$

where $\phi(x) = \max\{0, x - a\} = (x - a)_+$. Note that ϕ is convex and $\mathbb{E}|Y_k| < \infty$. Furthermore by Jensen and monotonicity of ϕ , (Y_n) is a submartingale:

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \geq \phi(\mathbb{E}[X_{n+1} \mid \mathcal{F}_n]) \geq \phi(X_n) = Y_n .$$

Let $V_{0,b-a}^{(n)}$ be the number of up-crossings for (Y_k) until time n on the interval $[0, b - a]$. By what we have already proved,

$$\mathbb{E}V_{0,b-a}^{(n)} \leq \frac{\mathbb{E}|Y_n|}{b-a} \leq \frac{\mathbb{E}|X_n| + |a|}{b-a} .$$

However up-crossings of (Y_n) on the interval $[0, b - a]$ exactly correspond to up-crossings of (X_n) on the interval $[a, b]$, so $V_{0,b-a}^{(n)} = U_{a,b}^{(n)}$, and the theorem is proved. $\square$

From this theorem it follows immediately that if (X_n) is bounded in L^1 then the number of up-crossings of any given interval is finite almost surely.

Corollary 9.4. *If (X_n) is a submartingale and $a < b$ then setting*

$$U_{a,b} = \lim_n U_{a,b}^{(n)} ,$$

the condition $\sup_n \mathbb{E}|X_n| < \infty$ implies $\mathbb{E}U_{a,b} < \infty$ (and therefore $U_{a,b} < \infty$ almost surely).

Proof. Since $U_{a,b}^{(n)}$ is non-decreasing in n , the monotone convergence theorem gives

$$\mathbb{E}U_{a,b} = \lim_n \mathbb{E}U_{a,b}^{(n)} \leq \sup_n \frac{\mathbb{E}|X_n| + |a|}{b-a} < \infty .$$

$\square$

LECTURE 17

9.1 Martingale convergence theorems

We now aim to show that L^1 bounded martingales and Doob martingales converge almost surely. We begin with L^1 bounded submartingales.

Theorem 9.5 (Martingale convergence). *Let (X_n) be a submartingale with $K = \sup_n \mathbb{E}|X_n| < \infty$. Then $X_n \rightarrow X$ almost surely for some random variable with $\mathbb{E}|X| \leq K$.*

Proof. This is a direct application of the up-crossing inequality. By the corollary of last section, for each a, b , $U_{a,b} < \infty$ almost surely. So

$$\mathbb{P}(U_{a,b} < \infty \text{ for all rational } a < b) = 1 .$$

Note that

$$\mathbb{P}(\liminf_n X_n < \limsup_n X_n) \leq \sum_{a < b \text{ rational}} \mathbb{P}\left(\liminf_n X_n < a < b < \limsup_n X_n\right) .$$

However if $\liminf_n X_n < a < b < \limsup_n X_n$ then $U_{a,b} = \infty$, so this has probability zero. We conclude that $\liminf_n X_n = \limsup_n X_n$ almost surely; that is, X_n converges to some X . By Fatou,

$$\mathbb{E}|X| \leq \liminf_n \mathbb{E}|X_n| \leq K .$$

□

Recall now that if X is an integrable random variable then given a filtration $(\mathcal{F}_n)$, a Doob martingale is defined as $X_n = \mathbb{E}[X | \mathcal{F}_n]$. We can identify the martingale limit in this case.

Theorem 9.6 (Convergence for Doob martingales). *Let X be integrable and $(\mathcal{F}_n)$ be a filtration. Writing $\mathcal{F} = \sigma(\cup_n \mathcal{F}_n)$,*

$$X_n \rightarrow \mathbb{E}[X | \mathcal{F}] \text{ almost surely .}$$

Proof. Because

$$\sup_n \mathbb{E}|X_n| = \sup_n \mathbb{E}|\mathbb{E}[X | \mathcal{F}_n]| \leq \mathbb{E}|X| < \infty ,$$

the martingale convergence theorem gives $X_n \rightarrow Z$ for some random variable Z with $\mathbb{E}|Z| \leq \mathbb{E}|X|$. We must show that $Z = \mathbb{E}[X | \mathcal{F}]$ almost surely and to do this we claim that

$$(X_n) \text{ is uniformly integrable .} \tag{13}$$

Assuming we show this, then take $A \in \mathcal{F}_m$ for some m and use uniform integrability to obtain

$$\int_A Z \, d\mathbb{P} = \lim_n \int_A X_n \, d\mathbb{P} = \lim_n \int_A \mathbb{E}[X | \mathcal{F}_n] \, d\mathbb{P} .$$

For $n \geq m$, the fact that $A \in \mathcal{F}_m \subset \mathcal{F}_n$ gives

$$\int_A Z \, d\mathbb{P} = \lim_n \int_A X \, d\mathbb{P} = \int_A X \, d\mathbb{P} .$$

These two integrals agree for A in the π -system $\cup_n \mathcal{F}_n$ and so they must agree in the generated sigma-algebra $\mathcal{F}$. This shows that $Z = \mathbb{E}[X \mid \mathcal{F}]$ almost surely.

To prove the claim, let $\epsilon > 0$. We must show that there exists $a > 0$ such that

$$\int_{\{|X_n| \geq a\}} |X_n| \, d\mathbb{P} < \epsilon \text{ for all } n .$$

By the definition of $X_n = \mathbb{E}[X \mid \mathcal{F}_n]$, we have $|X_n| \leq \mathbb{E}[|X| \mid \mathcal{F}_n]$ almost surely, so it will suffice to show that

$$\int_{\{|X_n| \geq a\}} \mathbb{E}[|X| \mid \mathcal{F}_n] \, d\mathbb{P} < \epsilon$$

and of course, since $\{|X_n| \geq a\} \in \mathcal{F}_n$, this is the same as

$$\int_{\{|X_n| \geq a\}} |X| \, d\mathbb{P} < \epsilon \text{ for all } n . \quad (14)$$

We now use a homework problem from last semester which states that since $X \in L^1$, we can find $\delta > 0$ such that for any $A \in \Sigma$ with $\mathbb{P}(A) < \delta$, we have

$$\int_A |X| \, d\mathbb{P} < \epsilon . \quad (15)$$

By Markov, for any $a > 0$,

$$\mathbb{P}(|X_n| \geq a) \leq \frac{1}{a} \mathbb{E}|X_n| \leq \frac{1}{a} \mathbb{E}[\mathbb{E}[|X| \mid \mathcal{F}_n]] = \frac{1}{a} \mathbb{E}|X| .$$

So choosing $a = \frac{2}{\delta \mathbb{E}|X|}$, this gives that $\mathbb{P}(|X_n| \geq a) < \delta$ and so (14) holds. $\square$

- In the case that $\mathcal{F} = \Sigma$; that is, the filtration generates the space, then

$$\mathbb{E}[X \mid \mathcal{F}_n] \rightarrow X \text{ almost surely .}$$

- If f_n is a dyadic martingale; that is, for some $f \in L^1([0, 1])$,

$$f_n(x) = 2^n \int_{\frac{k-1}{2^n}}^{\frac{k}{2^n}} f(t) \, dt \text{ for } x \in \left[\frac{k-1}{2^n}, \frac{k}{2^n} \right) ,$$

then $f_n \rightarrow f$ almost everywhere. (Note the similarity to the Lebesgue differentiation theorem.)

- For the branching process Z_n with offspring distribution that has mean m , we showed before that Z_n/m^n is a martingale. Note that it is also nonnegative, so

$$\sup_n \mathbb{E} \left| \frac{Z_n}{m^n} \right| = \sup_n \mathbb{E} Z_n / m^n = \mathbb{E} Z_1 / m = 1 ,$$

and Z_n/m^n converges almost surely to some variable Z with $\mathbb{E}|Z| \leq 1$. Since $m^n \rightarrow 0$ this means that $Z_n \rightarrow 0$. Since Z_n is integer valued, $Z_n = 0$ for all large n . In other words, if $m < 1$, the population dies out.

- If (X_n) is an L^1 bounded martingale that converges to X , we need not have $X_n = \mathbb{E}[X \mid \mathcal{F}_n]$. Indeed, define

$$f_n(x) = \begin{cases} 2^n & \text{if } x \in [0, 2^{-n}) \\ 0 & \text{if } x \in [2^{-n}, 1) \end{cases}$$

and let $\mathcal{F}_n$ be the sigma-algebra in $[0, 1)$ generated by the dyadic intervals of width 2^{-n} . Then you can check that (f_n) is a martingale that is nonnegative (and therefore is L^1 bounded) and converges to some f , which in this case is 0. However, $f_n \neq \mathbb{E}[0 \mid \mathcal{F}_n]$. The ingredient missing here is uniform integrability.

- If (X_n) is a uniformly integrable martingale, then there is a random variable $X \in L^1$ such that $X_n = \mathbb{E}[X \mid \mathcal{F}_n]$ and $X_n \rightarrow X$ almost surely and in L^1 . (In other words, both statements from the last two theorems hold.)

Proof. Because (X_n) is uniformly integrable, $\sup_n \mathbb{E}|X_n| < \infty$ and X_n converges almost surely to some X . So for any m and $A \in \mathcal{F}_m$, uniform integrability gives

$$\int_A X \, d\mathbb{P} = \lim_n \int_A X_n \, d\mathbb{P}.$$

However for $n \geq m$, $X_m = \mathbb{E}[X_n \mid \mathcal{F}_m]$, so the last integral is $\int_A X_m \, d\mathbb{P}$. This proves that $X_m = \mathbb{E}[X \mid \mathcal{F}_m]$. $\square$

LECTURE 18

10 Exchangeability

As an application of martingales, we will prove de Finetti's theorem on exchangeable sequences.

Definition 10.1. A sequence (X_n) of random variables is exchangeable if for each $n \geq 1$ and each permutation $\sigma \in S_n$,

$$(X_1, \dots, X_n) = (X_{\sigma(1)}, \dots, X_{\sigma(n)}) \text{ in distribution.}$$

The variables are exchangeable in the sense that we can permute (exchange) any finite set of them and leave the distributions unchanged. Here are some examples.

Examples.

1. If (X_n) is i.i.d. then it is exchangeable.

2. Consider an urn with one white ball and one red ball. At each time we remove a ball from the urn and return it with an additional ball of the same color. At time n , let

$$X_n = \begin{cases} 0 & \text{we draw a white ball} \\ 1 & \text{we draw a red ball} \end{cases}.$$

We claim the sequence (X_n) is exchangeable. To see why, let $n \geq 1$ and let $x_1, \dots, x_n \in \{0, 1\}$. Then

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n \frac{R_{i-1} \mathbf{1}_{\{x_i=1\}} + W_{i-1} \mathbf{1}_{\{x_i=0\}}}{R_{i-1} + W_{i-1} + 2},$$

where R_n is the number of red balls and W_n is the number of white balls at time n (with $R_0 = W_0 = 0$). The denominator above is $i + 1$. The numerator is either R_{i-1} or W_{i-1} depending on whether or not $x_i = 1$. Because we replace one ball at a time, the R_i 's for which $x_i = 1$ increase by one at a time. Therefore

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \frac{R_n! W_n!}{(n+1)n \cdots 2},$$

and this only depends on the numbers R_n and W_n . (Here $R_n = \sum_{i=1}^n x_i$.) For any permutation σ , the number $\sum_{i=1}^n x_{\sigma(i)}$ is the same as $\sum_{i=1}^n x_i$, so the above probability equals $\mathbb{P}(X_1 = x_{\sigma(1)}, \dots, X_n = x_{\sigma(n)})$.

3. Here is a more complicated example. Let ξ be a random variable supported on $[0, 1]$ and, conditioned on the value of ξ , let $(X_1, \dots)$ be an i.i.d. sequence taking values in $\{0, 1\}$ with probabilities $1 - \xi$ and ξ . We can construct this as follows. Let $Y_1, Y_2, \dots$ be i.i.d. uniform $[0, 1]$ random variables and define $(X_1, X_2, \dots)$ as a function of the sequence $(\xi, Y_1, Y_2, \dots)$ as

$$X_n = \mathbf{1}_{\{Y_n \leq \xi\}}.$$

We obtain

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n \mid \xi) = \xi^{\sum_i x_i} (1 - \xi)^{n - \sum_i x_i},$$

and because this only depends on $\sum_i x_i$, the sequence is exchangeable. Here we say that the sequence is a mixture of Bernoulli sequences.

In the last example, there is a random variable ξ such that, conditional on ξ , the X_n 's are i.i.d. Bernoulli. de Finetti's theorem says that this is always true.

Theorem 10.2. *Let (X_n) be an exchangeable sequence of $\{0, 1\}$ -valued random variables. There exists a random variable ξ such that for all $n \geq 1$ and $x_1, \dots, x_n \in \{0, 1\}$,*

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n \mid \xi) = \xi^s (1 - \xi)^{n-s}$$

and

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{E} \xi^s (1 - \xi)^{n-s},$$

where $s = \sum_{i=1}^n x_i$.

Proof. Note that the second equation follows from the first by integrating, so we need only prove the first.

In the construction above of i.i.d. mixtures, we see that (by the law of large numbers) ξ is actually the asymptotic fraction of 1's in the sequence (X_n) . So we will work to condition on this. We will do it by taking a large n , conditioning on the fraction of 1's (or number of 1's) and finding the conditional distribution of the first m variables, where m is fixed relative to n .

Define $S_n = X_1 + \cdots + X_n$. For a fixed value of $t = S_n$,

$$\mathbb{P}(S_n = t) = \sum_{x_1, \dots, x_n: \sum x_i = t} \mathbb{P}(X_1 = x_1, \dots, X_n = x_n) .$$

There are $\binom{n}{t}$ number of choices for the x_i 's and they are all equally likely. Therefore for each such choice (that is, each n -tuple $(x_1, \dots, x_n)$ such that $\sum_i x_i = t$),

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n \mid S_n = t) = \frac{1}{\binom{n}{t}} .$$

For $m \leq n$ and $x_1, \dots, x_m$ fixed such that $\sum_{i=1}^m x_i = s$, we can sum over all the $\binom{n-m}{t-s}$ possible entries $x_{m+1}, \dots, x_n$ for

$$\begin{aligned} \mathbb{P}(X_1 = x_1, \dots, X_m = x_m \mid S_n = t) &= \frac{\binom{n-m}{t-s}}{\binom{n}{t}} \\ &= \frac{(n-m)!t!(n-t)!}{(t-s)!(n-m-t+s)!n!} \\ &= \frac{[t]_s [n-t]_{m-s}}{[n]_m} , \end{aligned}$$

where $[a]_b = a!/(a-b)!$. Divide the top and bottom by n^m and rewrite this as

$$\frac{\frac{[t]_s}{n^s} \frac{[n-t]_{m-s}}{n^{m-s}}}{\frac{[n]_m}{n^m}} .$$

Setting $x = t/n$, the first term is

$$\begin{aligned} \frac{[t]_s}{n^s} &= \frac{t(t-1)\cdots(t-s+1)}{n^s} = \frac{t}{n} \left(\frac{t}{n} - \frac{1}{n} \right) \cdots \left(\frac{t}{n} - \frac{s-1}{n} \right) \\ &= x(x-1/n)\cdots(x-(s-1)/n) \\ &= \prod_{i=0}^{s-1} \left(x - \frac{i}{n} \right) . \end{aligned}$$

Similarly,

$$\frac{[n-t]_{m-s}}{n^{m-s}} = \prod_{i=0}^{m-s-1} \left(1 - x - \frac{i}{n}\right)$$

$$\frac{[n]_m}{n^m} = \prod_{i=0}^{m-1} \left(1 - \frac{i}{n}\right) .$$

Therefore

$$\mathbb{P}(X_1 = x_1, \dots, X_m = x_m \mid S_n = t) = f_{m,s,n}(t/n) , \quad (16)$$

where

$$f_{m,s,n}(x) = \frac{\prod_{i=0}^{s-1} \left(x - \frac{i}{n}\right) \prod_{i=0}^{m-s-1} \left(1 - x - \frac{i}{n}\right)}{\prod_{i=0}^{m-1} \left(1 - \frac{i}{n}\right)} .$$

We will finish next time. □

LECTURE 19

We were in the middle of proving:

Theorem 10.3. *Let (X_n) be an exchangeable sequence of $\{0, 1\}$ -valued random variables. There exists a random variable ξ such that for all $n \geq 1$ and $x_1, \dots, x_n \in \{0, 1\}$,*

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n \mid \xi) = \xi^s (1 - \xi)^{n-s}$$

and

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{E} \xi^s (1 - \xi)^{n-s} ,$$

where $s = \sum_{i=1}^n x_i$.

Proof. We showed that for $m \leq n$ and given s, t ,

$$\mathbb{P}(X_1 = x_1, \dots, X_m = x_m \mid S_n = t) = f_{m,s,n}(t/n) , \quad (17)$$

where

$$f_{m,s,n}(x) = \frac{\prod_{i=0}^{s-1} \left(x - \frac{i}{n}\right) \prod_{i=0}^{m-s-1} \left(1 - x - \frac{i}{n}\right)}{\prod_{i=0}^{m-1} \left(1 - \frac{i}{n}\right)} .$$

Because t takes only finitely many values of t in equation (17), we may rewrite it as

$$\mathbb{P}(X_1 = x_1, \dots, X_m = x_m \mid S_n) = f_{m,s,n} \left(\frac{S_n}{n} \right) .$$

Note that the entire development above only depended on choosing the value of $S_n = t$. In particular, if we also let $j \geq 1$ and fix values $t_1, \dots, t_j$ then

$$\mathbb{P}(X_1 = x_1, \dots, X_m = x_m \mid S_n = t, S_{n+1} = t_1, \dots, S_{n+j} = t_j) = f_{m,s,n}(t/n) ,$$

so long as the event we condition on has positive probability. In other words,

$$\mathbb{P}(X_1 = x_1, \dots, X_m = x_m \mid S_n, \dots, S_{n+j}) = f_{m,s,n} \left(\frac{S_n}{n} \right) .$$

Now we use martingale convergence. Because $\mathbf{1}_{\{X_1=x_1, \dots, X_m=x_m\}}$ is integrable, we can take a limit in j for

$$\mathbb{P}(X_1 = x_1, \dots, X_m = x_m \mid \mathcal{T}_n) = f_{m,s,n} \left(\frac{S_n}{n} \right) , \quad (18)$$

where $\mathcal{T}_n = \sigma(S_n, S_{n+1}, \dots)$. As n grows, the left side is a *backward martingale*.

Lemma 10.4. *Let X be integrable and $\mathcal{F}_1 \supset \mathcal{F}_2 \supset \dots$ be a nested sequence of sub-sigma-algebras of X . Then*

$$\mathbb{E}[X \mid \mathcal{F}_n] \rightarrow \mathbb{E}[X \mid \mathcal{F}] ,$$

where $\mathcal{F} = \cap_n \mathcal{F}_n$.

Proof. For any n , the sequence

$$X_n, X_{n-1}, \dots, X_1$$

is a martingale that is uniformly integrable (thus bounded in L^1). Therefore if we pick any rational $a < b$, we can still define the number $U_{a,b}^{(n)}$ of up-crossings of $[a, b]$ and $\mathbb{E}U_{a,b}^{(n)}$ is bounded. Furthermore it is still true that $U_{a,b}^{(n)}$ is non-decreasing in n , so by monotone convergence, $\mathbb{E}U_{a,b}$ is bounded, where $U_{a,b} = \lim_n U_{a,b}^{(n)}$. This implies as before that $X_n \rightarrow Z$ for some integrable Z . Note that Z is measurable relative to each $\mathcal{F}_n$, since it is a limit of X_n , which is measurable relative to $\mathcal{F}_n$. So Z is measurable relative to $\mathcal{F}$.

Now the proof that $Z = \mathbb{E}[X \mid \mathcal{F}]$ is nearly identical to what we showed for (forward) martingales under uniform integrability. $\square$

Applying the lemma in (18) and taking $n \rightarrow \infty$ we obtain

$$\mathbb{P}(X_1 = x_1, \dots, X_m = x_m \mid \mathcal{T}) = \lim_n f_{m,s,n} \left(\frac{S_n}{n} \right)$$

(and we are guaranteed the limit on the right exists, since the one on the left did). The formula for $f_{m,n,s}$ implies that if $x_n \rightarrow x$ then

$$f_{m,n,s}(x_n) \rightarrow x^s(1-x)^{m-s} .$$

So for a particular ω for which $\lim_n f_{m,s,n} \left(\frac{S_n}{n} \right)$ exists, if we let ξ be any limit point of $\left(\frac{S_n}{n} : n \geq m \right)$,

$$\mathbb{P}(X_1 = x_1, \dots, X_m = x_m \mid \mathcal{T}) = \xi^s(1-\xi)^{m-s} .$$

If we argue that there is only one limit point ξ then, since it is $\mathcal{T}$ measurable, it is a random variable (that is, it is Σ -measurable) and we will be done.

If $\xi_1(\omega) < \xi_2(\omega)$ are two different limit points of $\left(\frac{S_n}{n}(\omega) : n \geq m \right)$, then we must have

$$\xi_1^s(1-\xi_1)^{m-s} = \xi_2^s(1-\xi_2)^{m-s} ,$$

since they both equal the conditional probability. But since S_n has increments of size at most 1, each number in the interval $[\xi_1, \xi_2]$ must also be a limit point and satisfy this equation. Therefore the function $x \mapsto x^s(1-x)^{m-s}$ must be constant for a non degenerate interval of x 's and this is a contradiction. $\square$

11 Stochastic processes

A stochastic process is simply a collection $(X_t)_{t \in T}$ of random variables defined on the same space, indexed by some set T . We have already worked quite a bit with $T = \mathbb{N}$ (sequences of random variables) but also with $T = [0, \infty)$, when we studied the Poisson process. We will typically think of $T = [0, \infty)$ in what follows. The major issue here is that T can be uncountable, and we are not so well-equipped with dealing with uncountable collections.

Definition 11.1. *The finite dimensional distributions of a stochastic process (X_t) is the collection of distributions*

$$\{\mu_{t_1, \dots, t_k} : t_1, \dots, t_k \in T, k \geq 1\},$$

where $\mu_{t_1, \dots, t_k}$ is the distribution of the vector $(X_{t_1}, \dots, X_{t_k})$.

We note two properties of the finite dimensional distributions.

- If $k \geq 1$, $t_1, \dots, t_k \in T$ and $\sigma \in S_k$ is a permutation on k letters then for Borel sets $H_1, \dots, H_k \subset \mathbb{R}$,

$$\mu_{t_1, \dots, t_k}(H_1 \times \dots \times H_k) = \mu_{t_{\sigma(1)}, \dots, t_{\sigma(k)}}(H_{\sigma(1)} \times \dots \times H_{\sigma(k)}). \quad (19)$$

This is because each side equals $\mathbb{P}(X_{t_i} \in H_i \text{ for } i = 1, \dots, k)$.

- (Consistency) For $k \geq 1$, $t_1, \dots, t_{k+1} \in T$, then for Borel sets $H_1, \dots, H_k$,

$$\mu_{t_1, \dots, t_k}(H_1 \times \dots \times H_k) = \mu_{t_1, \dots, t_{k+1}}(H_1 \times \dots \times H_k \times \mathbb{R}). \quad (20)$$

This is because each side equals $\mathbb{P}(X_{t_i} \in H_i \text{ for } i = 1, \dots, k \text{ and } X_{t_{k+1}} \in \mathbb{R})$.

Now we would like to construct stochastic processes with given finite dimensional distributions. So given a collection $\mathbf{M} = \{\mu_{t_1, \dots, t_k} : t_1, \dots, t_k \in T, k \geq 1\}$ of measures we ask if there is a stochastic process with $\mathbf{M}$ as its finite dimensional distributions. Clearly (19) and (20) must hold. The theorem of Kolmogorov says that they are sufficient.

We will prove this theorem, but first let us talk a bit about product spaces again, because they will be essential for the construction.

LECTURE 20

Midterm will be a take home in place of the next homework.

11.1 Product spaces

Definition 11.2. Given a set T , the product space $\mathbb{R}^T$ is the set of “ T -dimensional vectors”

$$\mathbb{R}^T = \{x = (x_t)_{t \in T} : x_t \in \mathbb{R} \text{ for all } t\} .$$

This can be identified with the set of functions from T to $\mathbb{R}$.

Each element x consists of a collection of numbers $x_t = x(t)$, one for each $t \in T$. For each fixed $t \in T$ we can define the coordinate function

$$Z_t(x) = x_t = x(t) .$$

The coordinate function is a function of $x \in \mathbb{R}^T$ and for fixed t will correspond to an individual random variable X_t , so we would like it to be measurable. Therefore we define our product sigma-algebra when $\Omega = \mathbb{R}^T$ as

$$\Sigma = \mathcal{B}^T = \sigma(Z_t : t \in T) .$$

In other words, the sigma algebra is generated by the sets $\{x : x(t) \in H\}$ for Borel $H \subset \mathbb{R}$ and fixed $t \in T$.

Note that there are two points of view here. For a fixed x we may allow t to vary. This is thinking of x as a function of t and we can picture this as an arbitrary function from T to $\mathbb{R}$. For instance, when $T = [0, \infty)$, this is the collection of *all* functions from $[0, \infty)$ to $\mathbb{R}$ (however irregular). When we put a measure on the space, this point of view corresponds to thinking of random functions. On the other hand, we may fix t and view the coordinate map as a function of x . This point of view will correspond to thinking simply of a collection of random variables.

There is an alternative way to view Σ , as generated by the *finite dimensional cylinders*.

Definition 11.3. For $t_1, \dots, t_k \in T$ and Borel $H \in \mathbb{R}^k$, the k -dimensional cylinder corresponding to the t_i 's and H is

$$\{x \in \mathbb{R}^T : (Z_{t_1}(x), \dots, Z_{t_k}(x)) \in H\} .$$

Proposition 11.4. The collection Σ_0 consisting of all cylinder sets is an algebra.

Proof. We must show that $\Omega \in \Sigma_0$, that Σ is closed under complements and that it is closed under finite unions. For any t ,

$$\Omega = \{x : x_t \in \mathbb{R}\}$$

and if for $k \geq 1$, $t_1, \dots, t_k \in T$ with $H \subset \mathbb{R}^k$ Borel,

$$A = \{x : (x_{t_1}, \dots, x_{t_k}) \in H\}$$

then

$$A^c = \{x : (x_{t_1}, \dots, x_{t_k}) \in H^c\} ,$$

so the first two requirements are verified.

To show closure under unions, let A be as above and let B be

$$B = \{x : (x_{s_1}, \dots, x_{s_j}) \in I\}$$

for some $j \geq 1$, $s_1, \dots, s_j \in T$ and $I \subset \mathbb{R}^j$. We can rewrite A and B by just combining their index sets. This is conceptually pretty clear, but notationally it is not extremely fun. Enumerate

$$\{t_1, \dots, t_k\} \cup \{s_1, \dots, s_j\} = \{r_1, \dots, r_l\}$$

and defining two functions $\phi : \{1, \dots, k\} \rightarrow \{1, \dots, l\}$ and $\psi : \{1, \dots, j\} \rightarrow \{1, \dots, l\}$ by

$$\begin{aligned}\phi(i) &= m \text{ iff } t_i = r_m \\ \psi(i) &= m \text{ iff } s_i = r_m .\end{aligned}$$

Then we set

$$\begin{aligned}H' \subset \mathbb{R}^l &= \{(y_1, \dots, y_l) : (y_{\phi(1)}, \dots, y_{\phi(k)}) \in H\} \\ I' \subset \mathbb{R}^l &= \{(y_1, \dots, y_l) : (y_{\psi(1)}, \dots, y_{\psi(j)}) \in I\} .\end{aligned}$$

Now

$$A = \{x : (x_{r_1}, \dots, x_{r_l}) \in H'\} \text{ and } B = \{x : (x_{r_1}, \dots, x_{r_l}) \in I'\} ,$$

so that

$$A \cup B = \{x : (x_{r_1}, \dots, x_{r_l}) \in H' \cup I'\}$$

and $A \cup B \in \Sigma_0$. □

Now we can equivalently define the product sigma-algebra Σ by

$$\Sigma = \sigma(\Sigma_0) .$$

The advantage here is that it is now generated by an algebra, and we can use Caratheodory.

11.2 Kolmogorov consistency theorem

Theorem 11.5. *If $\mathbf{M}$ is a consistent family of finite dimensional distributions, indexed by k -tuples in T , then there exists a probability measure $\mathbb{P}$ on $\mathbb{R}^T$ such that*

$$\mu_{t_1, \dots, t_k}(\cdot) = \mathbb{P}(x : (x_{t_1}, \dots, x_{t_k}) \in \cdot) \text{ for all } \mu_{t_1, \dots, t_k} \in \mathbf{M} .$$

Proof. The strategy is to define $\mathbb{P}$ on Σ_0 and then extend to Σ uniquely using Caratheodory. So if A is a cylinder of the form

$$A = \{x : (x_{t_1}, \dots, x_{t_k}) \in H\} \text{ for } H \subset \mathbb{R}^k \text{ Borel ,}$$

define

$$\mathbb{P}(A) = \mu_{t_1, \dots, t_k}(H) .$$

Step 1. We will first show that $\mathbb{P}$ is well-defined. So suppose that A has two different representations: the one above and also

$$A = \{x : (x_{s_1}, \dots, x_{s_j}) \in I\} \text{ for } I \subset \mathbb{R}^j \text{ Borel} .$$

Then, as in the proof that Σ_0 is an algebra, we can find l and $H', I' \subset \mathbb{R}^l$ Borel such that

$$A = \{x : (x_{r_1}, \dots, x_{r_l}) \in H'\} = \{x : (x_{r_1}, \dots, x_{r_l}) \in I'\} .$$

Because we have moved from H to H' and I to I' by possibly adding coordinates and permuting them, we use the consistency conditions for

$$\mu_{t_1, \dots, t_k}(H) = \mu_{r_1, \dots, r_l}(H') \quad (21)$$

and

$$\mu_{s_1, \dots, s_j}(I) = \mu_{r_1, \dots, r_l}(I') . \quad (22)$$

We claim that in fact $H' = I'$. Indeed, if $(y_1, \dots, y_l) \in H' \subset \mathbb{R}^l$ then we can find a function $x \in \mathbb{R}^T$ such that $x(r_i) = y_i$ for $i = 1, \dots, l$ (just set x equal to these values at the r_i 's and, say, zero everywhere else). Then $x \in A$. But since A also equals $\{x : x_{r_1}, \dots, x_{r_l} \in I'\}$, we must have $(y_1, \dots, y_l) = (x_{r_1}, \dots, x_{r_l}) \in I'$. Thus $H' \subset I'$. Similarly $I' \subset H'$ and we have proved the claim.

Combine the fact that $H' = I'$ with (21) and (22) to obtain

$$\mu_{t_1, \dots, t_k}(H) = \mu_{s_1, \dots, s_j}(I) ;$$

that is, $\mathbb{P}(A)$ is well-defined.

We will continue next time. □

LECTURE 21

Last time we were proving:

Theorem 11.6. *If $\mathbf{M}$ is a consistent family of finite dimensional distributions, indexed by k -tuples in T , then there exists a probability measure $\mathbb{P}$ on $\mathbb{R}^T$ such that*

$$\mu_{t_1, \dots, t_k}(\cdot) = \mathbb{P}(x : (x_{t_1}, \dots, x_{t_k}) \in \cdot) \text{ for all } \mu_{t_1, \dots, t_k} \in \mathbf{M} .$$

Proof. If A is a cylinder of the form

$$A = \{x : (x_{t_1}, \dots, x_{t_k}) \in H\} \text{ for } H \subset \mathbb{R}^k \text{ Borel} ,$$

define

$$\mathbb{P}(A) = \mu_{t_1, \dots, t_k}(H) .$$

In step one, we showed that this is well-defined.

Step 2. $\mathbb{P}$ is a finitely additive measure on Σ_0 . First for any $t \in T$, $\mathbb{P}(\Omega) = \mu_t(\mathbb{R}) = 1$. Next if $A \in \Sigma_0$ is of the form $\{x : (x_{t_1}, \dots, x_{t_k}) \in H\}$ for $H \subset \mathbb{R}^k$ Borel, then

$$\mathbb{P}(A^c) = \mu_{t_1, \dots, t_k}(H^c) = 1 - \mu_{t_1, \dots, t_k}(H) = 1 - \mathbb{P}(A) .$$

Last if A, B are disjoint elements of Σ_0 then, as above, we can choose their index sets to be the same and write them as $A = \{x : (x_{t_1}, \dots, x_{t_k}) \in H\}$ and $B = \{x : (x_{t_1}, \dots, x_{t_k}) \in I\}$ for $H, I \subset \mathbb{R}^k$ Borel. Since A and B are disjoint, so must be H and I , so

$$\mathbb{P}(A \cup B) = \mu_{t_1, \dots, t_k}(H \cup I) = \mu_{t_1, \dots, t_k}(H) + \mu_{t_1, \dots, t_k}(I) = \mathbb{P}(A) + \mathbb{P}(B) .$$

Step 3. $\mathbb{P}$ is countably additive on Σ_0 . For this we must show that if $A_1 \supset A_2 \supset \dots \in \Sigma_0$ have $\cap_n A_n \in \Sigma_0$ then $\mathbb{P}(\cap_n A_n) = \lim_n \mathbb{P}(A_n)$. It suffices to consider the case that $\cap_n A_n = \emptyset$ and show that $\mathbb{P}(A_n) \rightarrow 0$.

Since (A_n) is nested decreasing, the fact that $\mathbb{P}$ is finitely additive on Σ_0 implies that $\mathbb{P}(A_n)$ is non-increasing in n so, by way of contradiction, assume that for some $\epsilon > 0$, $\mathbb{P}(A_n) \geq \epsilon$ for all n . Each A_n is of the form $\{x : (x_{t_1}, \dots, x_{t_{k_n}}) \in H_n\}$ for some Borel set $H_n \subset \mathbb{R}^{k_n}$. In the specification of A_n , we may always expand the index set $\{t_1, \dots, t_{k_n}\}$ by appending extra coordinates (and adding copies of $\mathbb{R}$ accordingly), so we may assume that the index sets are nested. Furthermore, by possibly repeating each A_n some number of times and adding Ω some number of times to the beginning of the sequence we may assume that $k_n = n$. Therefore we can find a sequence of indices $t_1, t_2, \dots$ and a sequence of Borel sets $H_n \subset \mathbb{R}^n$ such that

$$A_n = \{x : (x_{t_1}, \dots, x_{t_n}) \in H_n\} .$$

We will aim to get a contradiction by showing that $\cap_n A_n \neq \emptyset$. For this we will need to replace (A_n) by nested compact sets. But we need the following lemma, which is pretty standard in a real analysis course. We omit the proof, but you can find it in Billingsley, Theorem 12.3.

Lemma 11.7. (Regularity) *Let μ be a probability measure on $\mathbb{R}^n$. If $A \subset \mathbb{R}^n$ is a Borel set, then given $\epsilon > 0$, there exists $K \subset A$ that is compact and $\mu(A \setminus K) < \epsilon$.*

Using regularity, let $K_n \subset H_n$ be compact and such that

$$\mu_{t_1, \dots, t_n}(H_n \setminus K_n) < \epsilon/2^{n+1} .$$

Putting $B_n = \{x : (x_{t_1}, \dots, x_{t_n}) \in K_n\}$, then by definition of $\mu_{t_1, \dots, t_n}$,

$$\mathbb{P}(A_n \setminus B_n) < \epsilon/2^{n+1} .$$

We want the B_n 's to be nested, so define $C_n = B_1 \cap \dots \cap B_n$. Now

$$C_n \subset B_n \subset A_n$$

and because the A_k 's are nested,

$$\mathbb{P}(A_n \setminus C_n) \leq \sum_{k=1}^n \mathbb{P}(A_n \setminus B_k) \leq \sum_{k=1}^n \mathbb{P}(A_k \setminus B_k) < \sum_{k=1}^n \epsilon/2^{n+1} < \epsilon/2 .$$

This means that for all n ,

$$\epsilon < \mathbb{P}(A_n) \leq \mathbb{P}(A_n \setminus C_n) + \mathbb{P}(C_n) < \epsilon/2 + \mathbb{P}(C_n) ,$$

giving

$$\mathbb{P}(C_n) > \epsilon/2 \text{ for all } n .$$

We can now complete the proof. We have gotten ourselves into a situation where the C_n 's are compact and nested decreasing with $\mathbb{P}(C_n) > \epsilon/2$ for all n . This suggests the intersection should be nonempty. If we can show this, we will be done since $C_n \subset A_n$ and $\cap_n A_n = \emptyset$.

Since C_n has positive measure, it is nonempty. So we can pick $x^{(n)} \in C_n$. Looking at the k -th coordinate (for $k \geq 1$) and varying n , gives us the sequence

$$\{x_{t_k}^{(1)}, x_{t_k}^{(2)}, \dots\} . \quad (23)$$

For $n \geq k$, $x^{(n)} \in C_n \subset B_k$, so for such n , $(x_{t_1}^{(n)}, \dots, x_{t_k}^{(n)}) \in K_k$, which is a compact set. This means the sequence (23) is bounded and has a convergence subsequence. In fact, by a diagonal argument, we can find a sequence $n_1, n_2, \dots$ of integers such that

$$\lim_i x_{t_k}^{(n_i)} \text{ exists for each } k \geq 1 .$$

Calling this limit x_k , then each coordinate of $(x_{t_1}^{(n_i)}, x_{t_2}^{(n_i)}, \dots)$ converges as $i \rightarrow \infty$ to the corresponding coordinate of $x := (x_1, x_2, \dots)$.

Said differently, for fixed k , the vector $(x_1, \dots, x_k)$ is the limit (in i) of $(x_{t_1}^{(n_i)}, \dots, x_{t_k}^{(n_i)})$, a sequence in K_k , a compact set. Therefore $(x_1, \dots, x_k) \in K_k$, meaning $x \in B_k$ for each k . Thus $x \in \cap_k A_k$ and we are done. $\square$

12 Brownian motion (Wiener process)

For quite a while now we will discuss one particular stochastic process. It will be a collection of variables (X_t) for $t \in [0, \infty)$ intended to model random motion described by Billingsley as a particle “suspended in a fluid bombarded by molecules in thermal motion.” Our particle will start at position 0 at time 0; that is, $X_0 = 0$ almost surely. Then, for a fixed ω , the map $t \mapsto X_t(\omega)$ will model the trajectory of our particle undergoing random motion. It should be a continuous trajectory, but will appear jagged and irregular. It will be nowhere differentiable, but statistically it will actually have nice properties.

We can construct the Wiener process using the Kolmogorov consistency theorem from last section. To do this, we will need to specify finite dimensional distributions – for $k \geq 1$, $t_1, \dots, t_k \in [0, \infty)$ we must define the probability measures $\mu_{t_1, \dots, t_k}$ on $\mathbb{R}^k$. First a formal definition:

Definition 12.1. *A Wiener process is a stochastic process (X_t) such that*

1. $X_0 = 0$ almost surely,

2. for $t_1 < t_2 < \dots < t_k$, the increments $X_{t_1}, X_{t_2} - X_{t_1}, \dots, X_{t_k} - X_{t_{k-1}}$ are independent and
3. for $s \leq t$, the increment $X_t - X_s$ has mean zero Gaussian distribution with variance $t - s$.

To show existence of a Wiener process, we define our finite dimensional distributions to be those of Gaussian vectors. That is, for $t_1 < t_2 < \dots < t_k$ define the measure $\mu_{t_1, \dots, t_k}$ to be the distribution of a k dimensional Gaussian vector $(Z_{t_1}, \dots, Z_{t_k})$ with mean 0 and covariances

$$\text{Cov } Z_{t_i} Z_{t_j} = \min\{t_i, t_j\} .$$

For such variables, if $1 < i < j \leq k$,

$$\mathbb{E}((Z_{t_i} - Z_{t_{i-1}})(Z_{t_j} - Z_{t_{j-1}})) = t_i - t_i - t_{i-1} + t_{i-1} = 0$$

and for $1 < i = j$,

$$\mathbb{E}((Z_{t_i} - Z_{t_{i-1}})(Z_{t_j} - Z_{t_{j-1}})) = t_i - t_{i-1} .$$

Finally for $1 = i = j$, $\mathbb{E}Z_{t_i}Z_{t_j} = t_i$. In other words, the vector $(Z_{t_1}, \dots, Z_{t_k})$ has Gaussian independent increments with variance $t_i - t_{i-1}$. So we can alternatively define this vector by letting $Y_1, \dots, Y_k$ independent Gaussian such that Y_i has variance $t_i - t_{i-1}$ for $i > 1$ and t_1 for $i = 1$ and define the Z_{t_i} 's by $Z_{t_j} = Y_1 + \dots + Y_j$.

LECTURE 22

Precisely, for $t_1 < \dots < t_k \in [0, \infty)$, $\mu_{t_1, \dots, t_k}$ is the measure with density (relative to Lebesgue measure)

$$f_{t_1, \dots, t_k}(x_1, \dots, x_k) = \prod_{i=1}^k \frac{1}{\sqrt{2\pi(t_i - t_{i-1})}} \exp \left[-\frac{(x_i - x_{i-1})^2}{2(t_i - t_{i-1})} \right], \quad t_0 = x_0 = 0 .$$

For any permutation $\sigma \in S_k$ we define $\mu_{t_{\sigma(1)}, \dots, t_{\sigma(k)}}$ to be the distribution of $(Z_{t_{\sigma(1)}}, \dots, Z_{t_{\sigma(k)}})$. This defines a family of finite dimensional distributions **M**.

Proposition 12.2. **M** is consistent.

Proof. By definition, **M** satisfies the permutation requirement of consistency. So we must show that if $t_1, \dots, t_{k+1} \in [0, \infty)$ then for $H_1, \dots, H_k \subset \mathbb{R}$ Borel,

$$\mu_{t_1, \dots, t_k}(H_1 \times \dots \times H_k) = \mu_{t_1, \dots, t_{k+1}}(H_1 \times \dots \times H_k \times \mathbb{R}) .$$

By permuting the coordinates, it suffices to show that if $t_1 < \dots < t_k$ and $i \in \{1, \dots, k\}$ then for Borel $H_1, \dots, H_{i-1}, H_{i+1}, \dots, H_k \subset \mathbb{R}$,

$$\mu_{t_1, \dots, t_k}(H_1 \times \dots \times H_{i-1} \times \mathbb{R} \times H_{i+1} \times \dots \times H_k) \tag{24}$$

$$= \mu_{t_1, \dots, t_{i-1}, t_{i+1}, \dots, t_k}(H_1 \times \dots \times H_{i-1} \times H_{i+1} \times \dots \times H_k) . \tag{25}$$

So let $Y_1, \dots, Y_k$ be independent Gaussians with mean zero and variances $t_1, t_2 - t_1, \dots, t_k - t_{k-1}$. Then $\mu_{t_1, \dots, t_k}$ is the distribution of the vector $(S_1, \dots, S_k)$, where $S_i = Y_1 + \dots + Y_i$. Therefore (24) is

$$\mathbb{P}(S_1 \in H_1, \dots, S_{i-1} \in H_{i-1}, S_{i+1} \in H_{i+1}, \dots, S_k \in H_k) .$$

However $Y_i + Y_{i+1}$ is normally distributed with variance $(t_i - t_{i-1}) + (t_{i+1} - t_i) = t_{i+1} - t_{i-1}$. Therefore the vector $(S_1, \dots, S_{i-1}, S_{i+1}, \dots, S_k)$ (of length $k - 1$) has the same distribution as $(T_1, \dots, T_{k-1})$, where

$$T_i = W_1 + \dots + W_i .$$

and the W_i 's are independent Gaussians with mean zero and variances

$$t_1, \dots, t_{i-1} - t_{i-2}, t_{i+1} - t_{i-1}, t_{i+2} - t_{i-1}, \dots, t_k - t_{k-1} .$$

But the distribution of $(T_1, \dots, T_{k-1})$ is just $\mu_{t_1, \dots, t_{i-1}, t_{i+1}, \dots, t_k}$, and this gives (25). $\square$

12.1 Lack of continuity

The previous argument, along with Kolmogorov's theorem, guarantees the existence of a stochastic process with the finite dimensional distributions of that of Brownian motion. We even have $\mu_0 = \delta_0$ – that is, $X_0 = 0$ almost surely – since the distribution of X_0 is that of a centered Gaussian with variance zero. However, Kolmogorov's consistency theorem gives us nothing toward continuous sample paths. In other words, it could be that almost every path $t \mapsto X_t(\omega)$ is discontinuous. We will have to work harder to show the existence of a process with this property also.

In fact, one difficulty is that the set $\{x : x(t) \text{ is continuous on } [0, \infty)\} \subset \mathbb{R}^{[0, \infty)}$ is not even measurable in our product sigma-algebra. This is a consequence of the following result.

Theorem 12.3. *Let T be some index set and let $(X_t)_{t \in T}$ be a stochastic process. Then*

1.

$$\sigma(X_t : t \in T) = \bigcup_{S \subset T \text{ countable}} \sigma(X_s : s \in S) .$$

2. For any $S \subset T$,

$$\sigma(X_s : s \in S) = \{ \{ \omega : (X_s(\omega)) \in H \} : H \in \mathcal{B}^S \} .$$

The point of the first item is that the right side is actually a sigma-algebra. we do not have to generate one. Merely taking the union gives us a sigma-algebra. Why is this important? If $A \in \sigma(X_t : t \in T)$ then actually $A \in \sigma(X_s : s \in S)$ for some countable subset $S \subset T$ and therefore is “determined” by countably many times s . In the case of $\Omega = \mathbb{R}^T$ and $X_t = Z_t(x)$, this says that $\mathcal{B}^T = \bigcup_{S \subset T} \mathcal{B}^S$ for countable S .

The second item is an analogue of what we have seen before: if $(X_1, \dots, X_k)$ is a random vector then $\sigma(X_1, \dots, X_k)$ equals the collection of all sets of the form $\{(X_1, \dots, X_k) \in H\}$ for $H \subset \mathbb{R}^k$ Borel.

Proof. We begin with the first item. We may assume $T \neq \emptyset$. Call the left side Σ and the right Σ' . If $A \in \Sigma'$, then $A \in \sigma(X_s : s \in S)$ for some countable S . But this is a sub-sigma-algebra of Σ , so $A \in \sigma(X_t : t \in T)$, giving the containment $\supset$.

For the containment $\subset$ we first must show that Σ' is a sigma-algebra. Choosing $S = \{t\}$ for some $t \in T$, $\Omega \in \sigma(X_t) \subset \Sigma'$. Next if $A \in \Sigma'$ then $A \in \sigma(X_s : s \in S)$ for some countable S , and since this is a sigma-algebra, $A^c \in \sigma(X_s : s \in S) \subset \Sigma'$. Last if $A_1, A_2, \dots \in \Sigma'$ then we can find index sets $S_1, S_2, \dots \subset T$ that are countable and such that $A_n \in \sigma(X_s : s \in S_n)$ for all n . Then taking $S = \cup_n S_n$,

$$\sigma(X_s : s \in S_n) \subset \sigma(X_s : s \in S) \text{ for all } n$$

and therefore $A_n \in \sigma(X_s : s \in S)$ for all n . This is a sigma-algebra, so $\cup_n A_n \in \sigma(X_s : s \in S)$ and since S is countable, this is in Σ' .

Now Σ' is a sigma-algebra relative to which each X_t is measurable (just take $S = \{t\}$ for a given $t \in T$). By minimality, then, $\Sigma \subset \Sigma'$.

For the second item, the right side is a sigma-algebra (this is not hard to check) and contains the cylinders, so it contains the left side. To prove the other containment, define the function $\Phi : \Omega \rightarrow \mathbb{R}^S$ by

$$\Phi(\omega) = (X_s(\omega))_{s \in S} .$$

For any cylinder $\{X_{s_1} \in H_1, \dots, X_{s_k} \in H_k\}$, the pre image is an element of $\sigma(X_s : s \in S)$. Therefore Φ is measurable from $\sigma(X_s : s \in S)$ to $\mathcal{B}^S$ (since we only have to check the measurability condition on a generating set). This implies that for any Borel set $H \in \mathcal{B}^S$,

$$\{\omega : (X_s(\omega))_{s \in S} \in H\} = \Phi^{-1}(H) \in \sigma(X_s : s \in S)$$

and this is the other containment. □

Corollary 12.4. *Taking $T = [0, \infty)$, the set of continuous functions is not in $\mathcal{B}^T$.*

Proof. Suppose that $A = \{x \in \mathbb{R}^T : x(t) \text{ is continuous}\}$ is in

$$\mathcal{B}^T = \sigma(Z_t(x) : t \in T) .$$

From item 1 of the last theorem, we can find a countable set $S \subset T$ such that

$$A \in \sigma(Z_s(x) : s \in S) \subset \mathcal{B}^S .$$

From the second part, we can find a Borel set $H \in \mathcal{B}^S$ such that

$$A = \{x \in \mathbb{R}^T : (x(s))_{s \in S} \in H\} .$$

Now take any continuous function x and let $t_0 \notin S$. Define y by

$$y(t) = \begin{cases} x(t) & \text{if } t \neq t_0 \\ 1 + x(t_0) & \text{if } t = t_0 \end{cases} ,$$

so that y is not continuous. However $x \in A$ and $(y(s))_{s \in S} = (x(s))_{s \in S} \in H$, so $y \in A$, a contradiction. □

Because of this corollary, we will not even be able to talk about the set of continuous functions in the Borel sigma-algebra, so we cannot show that the Wiener process as constructed above has $\mathbb{P}(t \mapsto X_t \text{ is continuous}) = 1$ (this string of symbols is meaningless!). We would have had this same problem with the Poisson process last semester if we had tried to construct it using Kolmogorov's consistency, since right continuity is not accessible in $\mathcal{B}^T$. But we got around this by defining it differently, using i.i.d. sequences, and we will get around it this time as well.

Although Kolmogorov's consistency actually gives us uniqueness on the product sigma-algebra, this is not enough. Given a measure $\mathbb{P}$ on the product sigma-algebra, we must choose a process with this distribution. One option is what is given to us by the theorem: the coordinate process $(Z_t(x) : t \in T)$, but this will not be a good choice. This issue is analogous when choosing a random variable with a given distribution – it does not matter what the random variable does on a zero-measure set. So we could try to modify the coordinate process by changing it on a zero measure set. However there are uncountably many t 's and each particular variable $X_t(\omega)$ could be modified on a zero measure set. So we have to choose these exceptional sets consistently. As in the case of defining conditional distributions, we will solve this by looking at rationals.

LECTURE 23

12.2 Continuous modification

The next theorem proves the existence of a (standard) Brownian motion.

Theorem 12.5. *There exists a process $(B_t)_{t \in [0, \infty)}$ such that*

1. *for $k \geq 1$ and $0 \leq t_1 < t_2 < \dots < t_k$, $(B_{t_1}, \dots, B_{t_k})$ is Gaussian distributed with mean zero and $\text{Cov}(B_{t_i}, B_{t_j}) = \min\{t_i, t_j\}$ and*
2. *for all ω , the map $t \mapsto B_t(\omega)$ is continuous with $B_0(\omega) = 0$.*

Condition 1 says that (B_t) is a Gaussian process: its finite dimensional distributions are Gaussian.

Proof. Let (X_t) be the process given by Kolmogorov's consistency theorem; it satisfies the first condition. We will modify it so that the new process satisfies the second condition. On a probability one set, we will keep the values of (X_t) at dyadic rationals and redefine at other points.

Step 1. Uniform continuity at the dyadic rationals. Let D be the set of nonnegative dyadic rationals:

$$D = \left\{ \frac{k}{2^n} : n \geq 1, k \geq 1 \right\}$$

and

$$M_n = \max_{k=0, \dots, n2^n} \sup_{r \in D \cap \left[\frac{k}{2^n}, \frac{k+2}{2^n} \right]} |X_r - X_{\frac{k}{2^n}}|.$$

Here we have split $[0, n + 2^{1-n}]$ into overlapping intervals of size $2/2^n$ and are taking the suprema of differences of X_t for dyadic t in these intervals. If we show that M_n is small, this will mean that $|X_t - X_s|$ is small when t and s are nearby dyadic rationals.

Claim 12.6.

$$\sum_n \mathbb{P}(M_n \geq 1/n) < \infty .$$

Proof. To estimate M_n , fix a point $t = k/2^n$ and $m \geq 0$. We first estimate differences when $r \in D$ is in the $m + n$ -th level of dyadics: by the Etamadi maximal inequality and Markov,

$$\begin{aligned} \mathbb{P}\left(\max_{i=1,\dots,2^{m+1}} |X_t - X_{t+i/2^{m+n}}| \geq 1/n\right) &\leq 3 \max_{i=1,\dots,2^{m+1}} \mathbb{P}(|X_t - X_{t+i/2^{m+n}}| \geq 1/(3n)) \\ &\leq \max_{i=1,\dots,2^{m+1}} 3 \cdot (3n)^4 \mathbb{E}(X_t - X_{t+i/2^{m+n}})^4 . \end{aligned}$$

The variable $X_t - X_{t+i/2^{m+n}}$ is Gaussian with mean 0 and variance $i/2^{m+n}$, so it has fourth moment $3 \frac{i^2}{2^{2(m+n)}}$. So the above is bounded by

$$9 \cdot (3n)^4 \frac{2^{2(m+1)}}{2^{2(m+n)}} = \frac{36 \cdot (3n)^4}{2^{2n}} .$$

The events above are increasing in m and converge to give us the bound

$$\mathbb{P}\left(\max_{r \in D \cap [k2^{-n}, (k+2)2^{-n}]} |X_t - X_r| \geq 1/n\right) \leq \frac{36 \cdot (3n)^4}{2^{2n}} .$$

Summing this bound over n gives

$$\sum_n \mathbb{P}(M_n \geq 1/n) \leq \sum_n n 2^n \frac{36 \cdot (3n)^4}{2^{2n}} < \infty .$$

□

Let

$$E = \{\omega : \{M_n \geq 1/n\} \text{ occurs finitely often}\} \cap \{X_0 = 0\} .$$

By Borel-Cantelli and the last claim (and the fact that the distribution of X_0 is δ_0), $\mathbb{P}(E) = 1$. We claim that for $\omega \in E$, for any $T > 0$, $r \mapsto X_r(\omega)$ is uniformly continuous on the rationals in $[0, T]$. To show this, let $\epsilon > 0$ and $T > 0$. Pick n_0 such that for $n \geq n_0$, $\omega \in \{M_n < 1/n\}$ and choose $\delta = 1/n$, where n is such that

$$n > \max\{T, 2/\epsilon, n_0\} .$$

Letting $q, r \in D \cap [0, T]$ with $|q - r| < \delta$, these two numbers must lie in an interval of the form $[2^{-n}, (k+2)2^{-n}]$ for some $k = 1, \dots, n2^n$. So by the triangle inequality

$$|X_q - X_r| \leq 2 \sup_{s \in D \cap [k2^{-n}, (k+2)2^{-n}]} |X_s - X_{k2^{-n}}| \leq 2M_n \leq 2/n < \epsilon .$$

By uniform continuity, we can extend our process to another: define the process (B_t) by

$$B_t(\omega) = \begin{cases} \lim_{r \rightarrow t, r \in D} X_r(\omega) & \text{if } \omega \in E \\ 0 & \text{otherwise} \end{cases} .$$

Then $t \mapsto B_t(\omega)$ is continuous for all ω with $B_0 = 0$ and agrees with $X_t(\omega)$ for $\omega \in E$ and $t \in D$.

Step 2. We are left to show that (B_t) has the correct finite dimensional distributions. Fix $0 \leq t_1 < \dots < t_k \in [0, \infty)$ and choose sequences of dyadic rationals $(r_{1,n}), (r_{2,n}), \dots, (r_{k,n})$ such that $r_{i,n} \rightarrow t_i$ for all $i = 1, \dots, k$. Then since $t \mapsto B_t(\omega)$ is continuous for all ω ,

$$(B_{r_{1,n}}, \dots, B_{r_{k,n}}) \rightarrow (B_{t_1}, \dots, B_{t_k}) \text{ as } n \rightarrow \infty \text{ surely} .$$

The vector on the left equals $(X_{r_{1,n}}, \dots, X_{r_{k,n}})$ and because the times in this vector are dyadic, we know its joint distribution – it is Gaussian with density given by

$$f_n(x_1, \dots, x_k) = \prod_{i=1}^k \frac{1}{\sqrt{2\pi(r_{i,n} - r_{i-1,n})}} \exp \left[-\frac{(x_i - x_{i-1})^2}{2(r_{i,n} - r_{i-1,n})} \right], \quad r_{0,n} = x_0 = 0 .$$

As $n \rightarrow \infty$, this density converges to that of a Gaussian $(G_1, \dots, G_k)$ with mean zero and covariances $\text{Cov } G_i G_j = \min\{t_i, t_j\}$. So by dominated convergence, this must be the density function of $(X_{t_1}, \dots, X_{t_k})$ and we are done. $\square$

In fact a similar argument shows:

Theorem 12.7 (Kolmogorov continuity criterion). *Let $(X_t)_{t \in [0, \infty)}$ be a stochastic process and suppose that for each $T > 0$ there are positive constants α, β, K such that*

$$\mathbb{E}|X_t - X_s|^\alpha \leq K|t - s|^{1+\beta} \text{ for } s, t \in [0, T] .$$

Then there is a continuous modification of (X_t) . In other words, there is a process $(\tilde{X}_t)$ such that

1. $\omega \mapsto \tilde{X}_t(\omega)$ is continuous for all ω and
2. for each $t \geq 0$, $\mathbb{P}(X_t = \tilde{X}_t) = 1$.

For the case of Brownian motion we took $\alpha = 4$, $\beta = 1$, $K = 3$.

LECTURE 24

13 Properties of Brownian motion

From now on, when we talk of a Brownian motion, we mean one with the correct finite-dimensional distributions and also continuous paths.

13.1 Basic properties

If (B_t) is a Brownian motion then $X_1, X_2, \dots$, given by

$$X_n = B_n - B_{n-1}$$

is an i.i.d. standard Gaussian sequence. We will use this to derive:

Proposition 13.1. *With probability one,*

$$\limsup_{t \rightarrow \infty} \frac{B_t}{\sqrt{t}} = \infty \text{ and } \liminf_{t \rightarrow \infty} \frac{B_t}{\sqrt{t}} = -\infty .$$

Proof. The fact that $B_n/\sqrt{n}$ is a standard gaussian gives

$$\mathbb{P}\left(\frac{B_n}{\sqrt{n}} \geq M\right) = \mathbb{P}(B_1 \geq M) > 0 .$$

Therefore

$$\mathbb{P}(B_n \geq M\sqrt{n} \text{ i.o.}) > 0 .$$

However the event $\{B_n \geq M\sqrt{n} \text{ i.o.}\}$ is invariant under finite permutations of the i.i.d. variables $X_1, X_2, \dots$. So the Hewitt-Savage 0-1 law (from homework) shows that $B_n \geq M\sqrt{n}$ i.o. with probability one. But then

$$\mathbb{P}(\limsup_{t \rightarrow \infty} B_t/\sqrt{t} = \infty) \geq \mathbb{P}(\limsup_{n \rightarrow \infty} B_n/\sqrt{n} = \infty) = \lim_{M \rightarrow \infty} \mathbb{P}(\limsup_{n \rightarrow \infty} B_n/\sqrt{n} \geq M) = 1 .$$

The same argument applied to $-B_n$ shows that $\liminf_{t \rightarrow \infty} B_t/\sqrt{t} = -\infty$ almost surely. $\square$

Now we come to two useful transformations.

1. (Scale invariance) For any $c > 0$, the process $W_t = \frac{B_{ct}}{\sqrt{c}}$ is a Brownian motion.

Proof. W_t has continuous paths and $W_0 = 0$. Also for $0 \leq t_1 < \dots < t_k$, $(W_{t_1}, \dots, W_{t_k})$ is a Gaussian vector with mean zero and covariances

$$\text{Cov}(W_{t_i}, W_{t_j}) = (1/c)\mathbb{E}B_{ct_i}B_{ct_j} = (1/c)\min\{ct_i, ct_j\} = \min\{t_i, t_j\} .$$

So it has the correct finite-dimensional distributions. $\square$

2. For any $s \geq 0$, the process $(W_t)_{t \geq 0} = (B_{t+s} - B_s)_{t \geq 0}$ is a Brownian motion.

Proof. Again this is continuous with $W_0 = 0$. The finite dimensional distributions are Gaussian and you can check they have the same covariance. $\square$

3. (Inversion) The process (W_t) , given by

$$W_t = \begin{cases} tB_{1/t} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$$

is a Brownian motion.

Proof. W_t is continuous for $t > 0$ and for $0 < t_1 < \dots < t_k$, $(W_{t_1}, \dots, W_{t_k})$ is Gaussian with mean zero and covariance

$$\text{Cov}(W_{t_i}, W_{t_j}) = t_i t_j \mathbb{E} B_{1/t_i} B_{1/t_j} = t_i t_j \min\{1/t_i, 1/t_j\} = \min\{t_i, t_j\}.$$

Also for any $t > 0$,

$$\text{Cov}(W_t, W_0) = \mathbb{E} W_t W_0 = 0 = \min\{0, t\}.$$

So we need only show that W_t is continuous at $t = 0$. To do this, we use the same strategy as in the continuity proof for Brownian motion.

Let $n \geq 1$ and $m \geq 1$. Then (W_t) has the same finite-dimensional distributions as (B_t) and we can estimate using Etemadi

$$\mathbb{P}\left(\max_{k=0, \dots, 2^m} |W_{k/2^{m+n}}| \geq 1/n\right) \leq 3 \max_{k=0, \dots, 2^m} \mathbb{P}(|B_{k/2^{m+n}}| \geq 1/(3n)).$$

The variable $B_{k/2^{m+n}}$ is Gaussian with mean zero and variance $k/2^{m+n}$ so we apply the fourth moment again with Markov for the upper bound

$$3 \max_{k=0, \dots, 2^m} (3n)^4 \frac{3k^2}{2^{2m+2n}} \leq 9 \frac{(3n)^4}{2^{2n}}.$$

Taking $m \rightarrow \infty$,

$$\sum_n \mathbb{P}\left(\max_{r \in D, r \leq 2^{-n}} |W_r| \geq 1/n\right) < \infty,$$

where D is the set of dyadic rationals. By continuity of W_t for $t > 0$, if $W_t \geq 2/n$ for some $t \in [0, 2^{-n}]$ then $W_r \geq 1/n$ for a dyadic rational $r \in [0, 2^{-n}]$, so

$$\sum_n \mathbb{P}\left(\max_{t \in [0, 2^{-n}]} |W_t| \geq 2/n\right) < \infty.$$

By Borel-Cantelli, $\max_{t \in [0, 2^{-n}]} |W_t| \geq 2/n$ only occurs for finitely many n almost surely, so W_t is continuous at 0. $\square$

4. As a corollary, we see $\lim_{t \rightarrow \infty} \frac{B_t}{t} = 0$.

Proof. $\lim_{t \rightarrow \infty} \frac{B_t}{t} = 0$ if and only if $\lim_{u \rightarrow 0} u B_{1/u} = 0$, which holds by continuity at zero of Brownian motion. $\square$

5. Also the positive zero set

$$Z = \{t > 0 : B_t = 0\}$$

has $\inf Z = 0$. This follows from inversion, the above proposition and continuity. Indeed, take $W_t = tB_{1/t}$, so that W_t is a Brownian motion. Then almost surely,

$$\limsup_{t \rightarrow \infty} W_t = \infty \text{ and } \liminf_{t \rightarrow \infty} W_t = -\infty .$$

By the intermediate value theorem, we can find a random sequence $t_1, t_2, \dots \in [0, \infty)$ with $t_k \rightarrow \infty$ such that $W_{t_k} = 0$ for all k . Then $B_{1/t_k} = 0$ for all k and $\inf Z = 0$.

LECTURE 25

13.2 Non-differentiability

Using the results of last section, almost surely,

almost surely, B_t is not differentiable at $t = 0$.

Proof. Letting $W_t = tB_{1/t}$ as above, the fact that this is a Brownian motion implies that almost surely there is a sequence $t_1, t_2, \dots$ such that $t_i \rightarrow \infty$ and $W_{t_i} \rightarrow \infty$. Then

$$\frac{B_{1/t_i} - B_0}{1/t_i} = W_{t_i} \rightarrow \infty .$$

Therefore $\limsup_{h \downarrow 0} \frac{B_h - B_0}{h} = \infty$ almost surely. Similarly, $\liminf_{h \downarrow 0} \frac{B_h - B_0}{h} = -\infty$. $\square$

A consequence is that for fixed $s \geq 0$, B_t is almost surely not differentiable at s . This is from the fact that $(B_t - B_s)_{t \geq s}$ is a Brownian motion. We can say even more, treating all real s simultaneously.

Theorem 13.2. *There is an event A with $\mathbb{P}(A) = 1$ such that for all $\omega \in A$, the function $B_t(\omega)$ is not differentiable at any $t \geq 0$.*

The difference here is in the order of quantifiers. We can already prove that for any $s \geq 0$, almost surely B_t is not differentiable at s . We want to prove the stronger statement: almost surely, B_t is not differentiable at any $s \geq 0$.

Proof. We will show that the upper and lower derivatives are nowhere finite. That is, defining

$$D^+(t) = \limsup_{h \downarrow 0} \frac{B_{t+h} - B_t}{h}, \quad D^-(t) = \liminf_{h \downarrow 0} \frac{B_{t+h} - B_t}{h},$$

and setting $E = \{\omega : D^+(t) < \infty, D^-(t) > -\infty \text{ for some } t \geq 0\}$, we will show that E is contained in a probability zero event. (Note the phrasing here – we are not claiming E has probability zero, since we do not know it is measurable.) This will prove that for all ω in a probability 1 set, B_t is nowhere differentiable.

Take any $\omega \in E$. Then for each t , we can find an integer $K = K(t, \omega) > 0$ such that

$$-K < D^-(t) \leq D^+(t) < K .$$

This means that for some $\delta = \delta(t, \omega, K) > 0$,

$$|B_t - B_s| \leq K|s - t| \text{ for all } s \text{ with } t \leq s < t + \delta .$$

Because $\delta > 0$, there is some n_0 such that if $n \geq n_0$, we can find a $k = 0, \dots, n2^n$ such that

$$\frac{k-1}{2^n} < t \leq \frac{k}{2^n} < \frac{k+1}{2^n} < \frac{k+2}{2^n} < \frac{k+3}{2^n} < t + \delta .$$

Then by the triangle inequality,

$$|B_{\frac{k+j}{2^n}} - B_{\frac{k+j+1}{2^n}}| \leq K \frac{8}{2^n} \text{ for } j = 0, 1, 2 .$$

So define the event $A_n(K)$ as

$$A_n(K) = \bigcup_{k=0}^{n2^n} \left\{ |B_{\frac{k+j}{2^n}} - B_{\frac{k+j+1}{2^n}}| \leq K 2^{-n} \text{ for } j = 1, 2, 3 \right\} .$$

We have argued that if $\omega \in E$ then for some K , ω is in $A_n(K)$ for some K and all large n . In other words,

$$E \subset \bigcup_{K \in \mathbb{N}} \liminf_n A_n(K) .$$

We will show the event on the right has probability zero.

In the definition of $A_n(K)$, the increments for $j = 1, 2, 3$ are independent and each have the distribution of a normal variable with mean 0 and variance 2^{-n} . By scaling, this is the same distribution as $2^{-n/2}$ times a standard normal. So

$$\mathbb{P}(A_n(K)) \leq n2^n \left(\mathbb{P}(N(0, 1) \leq K \frac{8}{2^{n/2}}) \right)^3 .$$

The normal density is bounded by one, so we get the upper bound $n2^n(8K/2^{n/2})^3 = n(8K)^3/2^{n/2}$. In other words,

$$\text{for each } K \in \mathbb{N}, \mathbb{P}(A_n(K)) \rightarrow 0 \text{ as } n \rightarrow \infty ,$$

and this implies $\liminf_n A_n(K)$ has probability zero. Summing over $K \in \mathbb{N}$ completes the proof. $\square$

13.3 Stopping times revisited

In this section we analyze stopping times for stochastic processes. This will be the continuous analogue of what we have discovered earlier when studying martingales. Throughout, let $(X_t)_{t \geq 0}$ be a stochastic process.

Definition 13.3. We say that a collection of sub-sigma algebras $(\mathcal{F}_t)_{t \geq 0}$ of Σ is a filtration if $\mathcal{F}_s \subset \mathcal{F}_t$ whenever $s \leq t$. A stochastic process $(X_t)_{t \geq 0}$ is adapted to $(\mathcal{F}_t)$ if for all t , X_t is $\mathcal{F}_t$ -measurable.

Here are some general remarks about filtrations. Because t is a continuous variable, things are more nuanced than in the discrete case.

- $(\mathcal{F}_t)$ is *right-continuous* if, for all $t \geq 0$,

$$\mathcal{F}_t^+ = \mathcal{F}_t ,$$

where $\mathcal{F}_t^+ = \bigcap_{n \geq 1} \mathcal{F}_{t+1/n}$.

- Typically one assumes that $(\mathcal{F}_t)$ is not only right-continuous, but that $\mathcal{F}_0$ contains all events in

$$\mathcal{F}_\infty = \sigma(\cup_{t \geq 0} \mathcal{F}_t)$$

which have probability zero.

- One says that a filtration $(\mathcal{F}_t)$ on $(\Omega, \Sigma, \mathbb{P})$ is *standard* if it is right-continuous and all probability zero events of $\mathcal{F}_\infty$ are contained in $\mathcal{F}_0$. This set of probability zero events can be quite large: one typically assumes that the underlying space is *complete*. That is, one assumes that for all $A \in \Sigma$ such that $A \subset B$ for an event $B \in \Sigma$ with $\mathbb{P}(B) = 0$, one has $A \in \Sigma$. It follows that $\mathbb{P}(A) = 0$. From this point on, we will only work on complete spaces (this is not a loss in generality because one can always complete the space.)

As before, we can define *stopping times*.

Definition 13.4. Let $\tau : \Omega \rightarrow [0, \infty]$ and $(\mathcal{F}_t)$ be a filtration of (Ω) .

1. τ is called a *stopping time* relative to $(\mathcal{F}_t)$ if $\{\tau \leq t\} \in \mathcal{F}_t$ for all $t \geq 0$.
2. τ is called an *optional time* relative to $(\mathcal{F}_t)$ if $\{\tau < t\} \in \mathcal{F}_t$ for all $t \geq 0$.

Proposition 13.5. Every stopping time is an optional time. Furthermore τ is an optional time relative to $(\mathcal{F}_t)$ if and only if it is a stopping time relative to $(\mathcal{F}_t^+)$. Therefore if $(\mathcal{F}_t)$ is right-continuous, stopping times and optional times are the same.

Proof. If τ is a stopping time relative to $(\mathcal{F}_t)$ then for any $t \geq 0$,

$$\{\tau < t\} = \cup_{n \geq 1} \{\tau \leq t - 1/n\} \in \mathcal{F}_t .$$

So τ is an optional time relative to $(\mathcal{F}_t)$.

Suppose that τ is an optional time relative to $(\mathcal{F}_t)$. Then for $t \geq 0$, we can write the set

$$\{\tau \leq t\} = \cap_{n \geq 1} \{\tau < t + 1/n\} \in \cap_{n \geq 1} \mathcal{F}_{t+1/n} = \mathcal{F}_t^+ .$$

Conversely, if τ is a stopping time relative to $(\mathcal{F}_t^+)$ and $t \geq 0$,

$$\{\tau < t\} = \cup_{n \geq 1} \{\tau \leq t - 1/n\} \in \cup_{n \geq 1} \mathcal{F}_{t-1/n}^+ .$$

But each of these sigma-algebras is in $\mathcal{F}_t$, so $\{\tau < t\} \in \mathcal{F}_t$. □

LECTURE 26

As before,

Definition 13.6. For a stopping time τ , the stopped sigma-algebra is

$$\mathcal{F}_\tau = \{A \in \Sigma : A \cap \{\tau \leq t\} \in \mathcal{F}_t \text{ for all } t \geq 0\} .$$

Examples.

1. For any constant t , $\tau = t$ is both a stopping time and an optional time and $\mathcal{F}_\tau = \mathcal{F}_t$.
2. As before, if $\tau_1 \leq \tau_2$ then $\mathcal{F}_{\tau_1} \subset \mathcal{F}_{\tau_2}$.
3. Also as before, any stopping time τ is $\mathcal{F}_\tau$ -measurable.
4. If $B \subset \mathbb{R}$ then the *hitting time* of B for (X_t) is

$$\tau_B = \inf\{t \geq 0 : X_t \in B\} .$$

If (X_t) is continuous for all (ω) and B is closed, then this is a stopping time relative to the filtration $(\mathcal{F}_t)$ given by $\mathcal{F}_t = \sigma(X_s : s \leq t)$.

Proof. We must show that for $t \geq 0$, $\{\tau \leq t\} \in \mathcal{F}_t$. The process (d_t) given by

$$d_t = \inf\{|X_t - x| : x \in B\}$$

is continuous, since (X_t) is. Also it is adapted to $(\mathcal{F}_t)$. We claim that

$$\tau_B > t \text{ if and only if } d_s > 0 \text{ for all } s \leq t .$$

Take $s \leq t$ and assume that $d_s = 0$. Since B is closed, $X_s \in B$ and so $\{r \geq 0 : X_r \in B\}$ contains s . This implies that $\tau_B \leq t$. Conversely, if $\tau_B \leq t$ then for each $n \geq 1$, the set $\{r \geq 0 : X_r \in B\}$ contains an element r_n with $r_n \in [0, t + 1/n]$. Take any sub-sequential limit r of (r_n) . Since $X_{r_n} = 0$ for each n , continuity of X_t gives $X_r = 0$. But then $r \in [0, t]$ and this means $d_s = 0$ for some $s \leq t$ (take $s = r$).

Since d_t is continuous, $d_s > 0$ for all $s \leq t$ if and only if there is some $n \geq 1$ such that $d_s > 1/n$ for all $s \leq t$. Again By continuity,

$$\{d_s > 0 \text{ for all } s \leq t\} = \bigcup_{n \geq 1} \bigcap_{s \leq t, s \in \mathbb{Q}} \{d_s > 1/n\} ,$$

and this is in $\mathcal{F}_t$. Therefore $\{\tau \leq t\} = \{\tau > t\}^c$ is in $\mathcal{F}_t$. □

5. If (X_t) is right-continuous for all ω and B is open, then τ_B is an optional time relative to $(\mathcal{F}_t)$. You will prove this on homework.

13.4 Strong Markov property

Let's return to (B_t) as a Brownian motion. The filtration associated to (B_t) is $(\mathcal{F}_t)$, where

$$\mathcal{F}_t = \sigma(B_s : s \leq t)$$

We actually want to have a right-continuous filtration so we will also consider the filtration $(\mathcal{F}_t^+)$, where as before

$$\mathcal{F}_t^+ = \cap_{n \geq 1} \mathcal{F}_{t+1/n}.$$

Proposition 13.7. *For every $s \geq 0$, the process W_t , given by*

$$(W_t)_{t \geq 0} = (B_{t+s} - B_s)_{t \geq 0}$$

is a Brownian motion independent of $\mathcal{F}_s^+$. In other words, $\sigma(W_t : t \geq 0)$ is independent of $\mathcal{F}_s^+$.

Proof. We have seen that (W_t) is a Brownian motion. First let us show that it is independent of $\mathcal{F}_s$. Let $k \geq 1$ and $0 \leq t_1 < \dots < t_k$. The set of events that are independent of $W = (W_{t_1}, \dots, W_{t_k})$ forms a λ -system, so we must check that it is independent of a π -system that generates $\mathcal{F}_s$. We will use the cylinders for $(B_r : r \leq s)$. So let $H \subset \mathbb{R}^k$ be Borel and $0 \leq r_1 < \dots < r_l \leq s$ with $I \subset \mathbb{R}^l$ Borel. Then

$$\begin{aligned} \mathbb{P}((W_{t_1}, \dots, W_{t_k}) \in H, (B_{r_1}, \dots, B_{r_l}) \in I) \\ = \mathbb{P}((W_{t_1}, \dots, W_{t_k}) \in H, (B_{r_1}, B_{r_2} - B_{r_1}, \dots, B_{r_l} - B_{r_{l-1}}) \in I'), \end{aligned}$$

where $I' = \{(x_1, \dots, x_l) : (x_1, x_1 + x_2, \dots, x_1 + \dots + x_l) \in I\}$. By independence of increments, this equals

$$\mathbb{P}((W_{t_1}, \dots, W_{t_k}) \in H) \mathbb{P}((B_{r_1}, B_{r_2} - B_{r_1}, \dots, B_{r_l} - B_{r_{l-1}}) \in I'),$$

which is $\mathbb{P}((W_{t_1}, \dots, W_{t_k}) \in H) \mathbb{P}((B_{r_1}, \dots, B_{r_l}) \in I)$. This shows $\{(W_{t_1}, \dots, W_{t_k}) \in H\}$ is independent of $\mathcal{F}_s$. The set of events independent of $\mathcal{F}_s$ is a λ -system and we have checked independence for a π -system generating $\sigma(W_t : t \geq 0)$.

Now that (W_t) is independent of $\mathcal{F}_s$, we show it is actually independent of $\mathcal{F}_s^+$. Let $N > 0$. The vector $(B_{t_1+s+1/N} - B_{s+1/N}, \dots, B_{t_k+s+1/N} - B_{s+1/N})$ is independent of $\mathcal{F}_{s+1/N}$ and therefore of $\mathcal{F}_{s+1/n}$ for $n \geq N$. Thus this vector is independent of $\mathcal{F}_s^+$. However by continuity of Brownian motion,

$$(B_{t_1+s+1/N} - B_{s+1/N}, \dots, B_{t_k+s+1/N} - B_{s+1/N}) \rightarrow (W_{t_1}, \dots, W_{t_k}) \text{ for all } \omega.$$

Therefore we can finish by the following lemma.

Lemma 13.8. *If (X_N) is a sequence of random vectors in $\mathbb{R}^k$ that are independent of a sigma-algebra $\mathcal{F}$ and $X_N \rightarrow X$ point wise then X is independent of $\mathcal{F}$.*

Proof. Let $A \in \mathcal{F}$ and $H \subset \mathbb{R}^k$ be Borel. If $\mathbb{P}(A) = 0$ then

$$\mathbb{P}(A, X_N \in H) = 0 = \mathbb{P}(A)\mathbb{P}(X_N \in H) .$$

Otherwise define the measure $\mathbb{P}_A(\cdot) = \mathbb{P}(\cdot \mid A)$. We want to show that the distribution of X under $\mathbb{P}_A$ is the same as the distribution of X under $\mathbb{P}$. This will show independence. So let $\phi_{A,X}$ be the characteristic function of X under $\mathbb{P}_A$ and ϕ_X the characteristic function of X under $\mathbb{P}$. Because $X_N \rightarrow X$ point wise, the dominated convergence theorem gives both $\phi_{X_N} \rightarrow \phi_X$ point wise and $\phi_{A,X_N} \rightarrow \phi_{A,X}$ point wise. However by independence, $\phi_{A,X_N} = \phi_{X_N}$ for all N and so $\phi_X = \phi_{A,X}$. $\square$

$\square$

LECTURE 27

Much of the material for the next few lectures is taken from *Brownian motion* by Peres and Mörters.

Corollary 13.9. *(B_t) is a continuous time martingale relative to $(\mathcal{F}_t^+)$. That is*

1. $\mathbb{E}|B_t| < \infty$ for all $t \geq 0$,
2. B_t is $\mathcal{F}_t^+$ measurable for all $t \geq 0$ and
3. if $s \leq t$ then $\mathbb{E}[B_t \mid \mathcal{F}_s^+] = B_s$ almost surely.

Proof. We need only check the last condition. Write

$$\begin{aligned} \mathbb{E}[B_t \mid \mathcal{F}_s] &= \mathbb{E}[B_s \mid \mathcal{F}_s^+] + \mathbb{E}[B_t - B_s \mid \mathcal{F}_s^+] \\ &= B_s + \mathbb{E}[B_t - B_s \mid \mathcal{F}_s^+] . \end{aligned}$$

However $B_t - B_s$ is independent of $\mathcal{F}_s^+$, so this is B_s . $\square$

We will from now consider the filtration $(\mathcal{F}_t^+)$. Since it is right-continuous, its stopping times and optional times are the same.

Theorem 13.10 (Strong Markov property). *Let τ be an almost surely finite stopping time and define*

$$W_t = B_{\tau+t} - B_\tau \text{ for } t \geq 0 .$$

Then $(W_t)_{t \geq 0}$ is a Brownian motion independent of $\mathcal{F}_\tau^+$.

In the statement we are using the stopped sigma-algebra relative to $(\mathcal{F}_t^+)$. It is

$$\mathcal{F}_\tau^+ = \{A \in \Sigma : A \cap \{\tau \leq t\} \in \mathcal{F}_t^+ \text{ for all } t \geq 0\} .$$

Proof. First approximate τ from above for a fixed $n \geq 1$ by

$$\tau_n = \frac{k+1}{2^n} \text{ iff } \tau \in \left[\frac{k}{2^n}, \frac{k+1}{2^n} \right), \quad k = 0, 1, \dots$$

Note that τ_n is almost surely finite and it is a stopping time since for t of the form $(k+1)/2^n$,

$$\{\tau_n \leq t\} = \{\tau < (k+1)2^{-n}\} \in \mathcal{F}_{(k+1)2^{-n}}^+ = \mathcal{F}_t^+.$$

The times τ_n decrease to τ and we will first show the statement for the times τ_n .

So set

$$B_t^{(k,n)} = B_{t+k/2^n} - B_{k/2^n} \text{ for } t \geq 0 \text{ and fixed } k \geq 0$$

and

$$W_t^{(n)} = B_{t+\tau_n} - B_t \text{ for } t \geq 0.$$

For each event of the type $\{W^{(n)} \in A\}$ for $A \subset \mathbb{R}^{[0,\infty)}$ Borel and $E \in \mathcal{F}_{\tau_n}^+$,

$$\mathbb{P}(W^{(n)} \in A, E) = \sum_{k=0}^{\infty} \mathbb{P}(B_k \in A, E, \tau_n = k/2^n).$$

The event $E \cap \{\tau_n = k/2^n\}$ is in $\mathcal{F}_{k/2^n}^+$ and by independence of B_k (which is just a Brownian motion) from this sigma-algebra, we obtain

$$\sum_{k=0}^{\infty} \mathbb{P}(B_k \in A) \mathbb{P}(E, \tau_n = k/2^n) = \mathbb{P}(B \in A) \sum_{k=0}^{\infty} \mathbb{P}(E, \tau_n = k/2^n) = \mathbb{P}(B \in A) \mathbb{P}(E).$$

Taking $E = \Omega$ we see that $W^{(n)}$ is a Brownian motion. Because this holds for all E and A , it also shows that it is independent of $\mathcal{F}_t^+$.

For the full statement, take $k \geq 1$, $0 \leq t_1 < \dots < t_k$ and $E \in \mathcal{F}_{\tau}^+$. Since $\tau \leq \tau_n$ for all n , $\mathcal{F}_{\tau}^+ \subset \mathcal{F}_{\tau_n}^+$ (you will show this on homework), so $E \in \mathcal{F}_{\tau_n}^+$. Therefore using the first case, $(W_{t_1}^{(n)}, \dots, W_{t_k}^{(n)})$ is independent of E . Since $\tau_n \downarrow \tau$ and B_t is continuous,

$$(W_{t_1}^{(n)}, \dots, W_{t_k}^{(n)}) \rightarrow (W_{t_1}, \dots, W_{t_k}) \text{ for all } \omega.$$

By the lemma last time, $(W_{t_1}, \dots, W_{t_k})$ is also independent of E , so this proves the independence assertion. Furthermore, since point wise convergence implies convergence in distribution, (W_t) is also a Brownian motion. $\square$

13.5 Applications of the strong Markov property

The Zero set.

For our first application we consider the zero set of Brownian motion:

$$Z = \{t \geq 0 : B_t = 0\}.$$

Theorem 13.11. *The set $Z = Z(\omega)$ almost surely has the following properties.*

1. Z is uncountable.
2. Z is nowhere dense.
3. Z is perfect (closed and every point is a limit point of it).
4. Z is unbounded and has Lebesgue measure zero.

Proof. Because B_t is continuous, $Z = B_t^{-1}\{0\}$ is closed. Also since $\sup_t B_t = \infty$ and $\inf_t B_t = -\infty$, continuity also gives Z unbounded almost surely. To show that Z is nowhere dense we must show that almost surely, Z is not dense in any interval. So assume that Z is dense in a non-empty interval I . Then the closure of Z (which is Z , since it is closed) contains I . But then B_t must be differentiable on this interval, a contradiction. Thus if Z is not nowhere dense, B_t must be differentiable on an interval, and this is a probability zero event.

Once we show Z is perfect, it must be uncountable (this is an exercise in Rudin, Ch. 2). So we are left to show this and that Z has Lebesgue measure zero. For the latter we need a lemma.

Lemma 13.12. *(B_t) is jointly measurable. That is, the map $(t, \omega) \mapsto B_t(\omega)$ from $[0, \infty) \times \Omega \rightarrow \mathbb{R}$ is measurable relative to the product sigma-algebra $\mathcal{B} \times \Sigma$ (where $\mathcal{B}$ is the Borel sigma-algebra on $[0, \infty)$).*

□

LECTURE 28

We were proving properties of the zero set of Brownian motion:

$$Z = \{t \geq 0 : B_t = 0\} .$$

Theorem 13.13. *The set $Z = Z(\omega)$ almost surely has the following properties.*

1. Z is uncountable.
2. Z is nowhere dense.
3. Z is perfect (closed and every point is a limit point of it).
4. Z is unbounded and has Lebesgue measure zero.

Proof. We finished last time stating the lemma:

Lemma 13.14. *(B_t) is jointly measurable. That is, the map $(t, \omega) \mapsto B_t(\omega)$ from $[0, \infty) \times \Omega \rightarrow \mathbb{R}$ is measurable relative to the product sigma-algebra $\mathcal{B} \times \Sigma$ (where $\mathcal{B}$ is the Borel sigma-algebra on $[0, \infty)$).*

Proof. We will exhibit B_t is a limit of jointly measurable processes. For each n , define

$$B_t^{(n)}(\omega) = B_{k/2^n}(\omega) \text{ iff } \frac{k}{2^n} \leq t < \frac{k+1}{2^n} .$$

Then for fixed (t, ω) , continuity of B_t gives $B_t^{(n)}(\omega) \rightarrow B_t(\omega)$, so we must show that $B_t^{(n)}(\omega)$ is jointly measurable for each n . So letting $H \subset \mathbb{R}$ be Borel, note that

$$\{(t, \omega) : B_t(\omega) \in H\} = \bigcup_{k \in \mathbb{Z}} \left(\left[\frac{k}{2^n}, \frac{k+1}{2^n} \right) \times \{\omega : B_{k/2^n}(\omega) \in H\} \right) .$$

The right side is in $\mathcal{B} \times \Sigma$ so we are done. $\square$

By the lemma, for any $T > 0$, we can apply Fubini to the set $\{(t, \omega) : B_t(\omega) = 0\}$; letting λ be Lebesgue measure on $[0, T]$,

$$\mathbb{E}[\lambda(Z(\omega))] = (\lambda \times \mathbb{P}) \{(t, \omega) : B_t(\omega) = 0, t \in [0, T]\} = \int_0^T \mathbb{P}(B_t = 0) d\lambda(t)$$

equals zero. Therefore $\lambda(Z) = 0$ almost surely.

Last we show that each point of Z is a limit point of Z . So let $z \in Z$. We already know that if $z = 0$ then z is almost surely a limit point of Z , so assume that $z > 0$. If there is a sequence of zeroes of B_t converging up to z then we are done, so we may assume that for some rational $r < z$, z is the first zero of B_t after r . In other words,

$$z = \inf\{t \geq r : B_t = 0\} .$$

Now for each rational $r > 0$, the time $\tau_r = \inf\{t \geq r : B_t = 0\}$ is a stopping time (the proof is a slight modification of the one that says that hitting times of closed sets are stopping times for continuous processes). So since τ_r is almost surely finite, the strong Markov property says that $(B_t^{(r)}) = (B_{t+\tau_r} - B_{\tau_r})$ is a Brownian motion. Since 0 is almost surely a limit point of the zero set, we may find an event E_r of probability one such that for $\omega \in E_r$, there is a sequence $r_1 > r_2 > \dots > \tau_r$ such that each $r_i \in Z(\omega)$.

Now set $E = \bigcap_{r \in \mathbb{Q}: r > 0} E_r$, so that E has probability one. We claim that for $\omega \in E$, $Z(\omega)$ has no isolated points. If $z \in Z(\omega)$ is isolated then $z > 0$ since $\tau_0 = 0$. We can then choose a rational $r > 0$ such that there is no zero of B_t between r and z . It follows that $\tau_r = z$ and since $\omega \in E_r$, z is not isolated (from above), giving a contradiction. $\square$

Reflection principle and the distribution of the maximum.

For the next application of strong Markov, we find the distribution of the maximum of B_t on an interval. We will use the *reflection principle*: if τ is an almost surely finite stopping time, we define a new process (B_t^*) which is B_t for $t \leq \tau$ and for $t > \tau$, it is B_t reflected about the value B_τ :

$$B_t^* = \begin{cases} B_t & \text{if } t \leq \tau \\ B_\tau - (B_t - B_\tau) & \text{if } t > \tau \end{cases} .$$

We will show soon that this is a Brownian motion. But first we use it to compute the distribution of

$$M_t := \max_{0 \leq s \leq t} B_s .$$

Let $a > 0$ and τ_a be the (almost surely finite) hitting time of $\{a\}$. Then for $t \geq 0$,

$$\begin{aligned} \mathbb{P}(M_t \geq a) &= \mathbb{P}(\tau_a \leq t) \\ &= \mathbb{P}(\tau_a \leq t, B_t \geq a) + \mathbb{P}(\tau_a \leq t, B_t < a) \\ &= \mathbb{P}(B_t \geq a) + \mathbb{P}(\tau_a \leq t, B_t^* > a) . \end{aligned}$$

In the last line we have used that if $B_t \geq a$ then $\tau_a \leq t$. Also we have used B_t^* as the reflected Brownian motion defined above, associated to the stopping time τ_a . If we define τ_a^* as the hitting time of $\{a\}$ by B_t^* , then note that whenever $B_t^* > a$, we must have $\tau_a = \tau_a^*$, so

$$\mathbb{P}(\tau_a \leq t, B_t^* > a) = \mathbb{P}(\tau_a^* \leq t, B_t^* > a) = \mathbb{P}(B_t^* > a) ,$$

and this equals $\mathbb{P}(B_t > a) = \mathbb{P}(B_t \geq a)$. Therefore

$$\mathbb{P}(M_t \geq a) = 2\mathbb{P}(B_t \geq a) = \frac{2}{\sqrt{2\pi t}} \int_a^\infty e^{-\frac{x^2}{2t}} dx .$$

Theorem 13.15 (Reflection principle). *If τ is an almost surely finite stopping time then the reflected process (B_t^*) is a Brownian motion.*

Proof. Reflection is a continuous operation, so the paths of B_t^* are still continuous. So we must only check the finite-dimensional distributions. Billingsley only checks the two-point distributions “for notational convenience.” So to do this, first define the approximate stopping time τ_n by

$$\tau_n = k/2^n \text{ if and only if } \tau \in [k/2^n, (k+1)/2^n) .$$

We will show that the two-point distributions of $(B_t^{(n)*})$, the Brownian motion reflected using τ_n , are those of Brownian motion. Since $B_t^{(n)*} \rightarrow B_t^*$ point wise, the two-point distributions of $(B_t^{(n)*})$ converge to those of (B_t^*) , and since they are those of Brownian motion, so too must be those of $(B_t^{(n)})$.

τ_n takes countably many values (only dyadic rationals), so let t_0 be one of them. Let $s \leq t$ be two times; we want to show that for $H \subset \mathbb{R}^2$ Borel,

$$\mathbb{P}(\tau_n = t_0, (B_s, B_t) \in H) = \mathbb{P}(\tau_n = t_0, (B_s^*, B_t^*) \in H) . \quad (26)$$

Then summing over t_0 we will be done. For $s \leq t \leq t_0$ this is true since then $B_s^* = B_s$ and $B_t^* = B_t$. For $s \leq t_0 \leq t$, we consider $I \subset \mathbb{R}^2$ and $J \subset \mathbb{R}$ Borel and get

$$\begin{aligned} \mathbb{P}(\tau_n = t_0, (B_s, B_{t_0}) \in I, B_t - B_{t_0} \in J) &= \mathbb{P}(\tau_n = t_0, (B_s, B_{t_0}) \in I, B_{t_0} - B_t \in J) \\ &= \mathbb{P}(\tau_n = t_0, (B_s^*, B_{t_0}^*) \in I) \mathbb{P}(B_{t_0}^* - B_t^* \in J) \\ &= \mathbb{P}(\tau_n = t_0, (B_s^*, B_{t_0}^*) \in I, B_{t_0}^* - B_t^* \in J) . \end{aligned}$$

So for $K = I \times J$,

$$\mathbb{P}(\tau_n = t_0, (B_s, B_{t_0}, B_t - B_{t_0}) \in K) = \mathbb{P}(\tau_n = t_0, (B_s^*, B_{t_0}^*, B_t^* - B_{t_0}^*) \in K) .$$

By the π - λ theorem, this holds for all Borel $K \subset \mathbb{R}^3$. So set

$$K = \{(x_1, x_2, x_3) : (x_1, x_2 + x_3) \in H\} ,$$

and we obtain (26).

Billingsley says that the case $t_0 \leq s \leq t$ is similar. We can consider the joint distribution of $(B_{t_0}, B_s - B_{t_0}, B_t - B_s)$ instead of $(B_s, B_{t_0}, B_t - B_{t_0})$. $\square$

14 Brownian motion in higher dimensions

Definition 14.1. Let $B^{(1)}, \dots, B^{(d)}$ be independent Brownian motions. Then

$$B(t) = (B_t^{(1)}, \dots, B_t^{(d)})$$

is called a (standard) d -dimensional Brownian motion.

There are some differences between one-dimensional and higher-dimensional Brownian motion, for instance with regard to recurrence.

LECTURE 29

For $x \in \mathbb{R}^d$, the process

$$(x + B_t)_{t \geq 0}$$

is a (d -dimensional) Brownian motion *started at* x . For notational convenience, let us place the dependence on x in the measure: let $\mathbb{P}_x$ be the probability measure for which (B_t) is a Brownian motion started at x .

Theorem 14.2. Brownian motion is point recurrent if and only if $d = 1$. In other words, for $x, y \in \mathbb{R}^d$,

$$\mathbb{P}_x(B_t = y \text{ for some } t \in (0, \infty]) = 0$$

if and only if $d \geq 2$.

The route to proving this theorem takes us first to another result. It says that if we view two-dimensional Brownian motion as a plane curve (from time 0 to 1) then this curve has area zero.

Theorem 14.3 (Area of planar Brownian motion). Let λ be two-dimensional Lebesgue measure and (B_t) a two-dimensional Brownian motion. Then

$$\lambda\{B_t : t \in [0, 1]\} = 0 .$$

Proof. The idea will be to show that almost surely, different increments of Brownian motion intersect each other in zero measure sets. This will be proved using a scaling trick. Then we will use a general lemma to show that this is impossible unless the increments have zero area themselves.

Let $X = \lambda\{B_t : t \in [0, 1]\}$. If X has finite mean, then since $\sqrt{3}B_t$ has the same distribution as B_{3t} , we find

$$3\mathbb{E}X = \mathbb{E}\lambda\{\sqrt{3}B_t : t \in [0, 1]\} = \mathbb{E}\lambda\{B_t : t \in [0, 3]\} \leq \sum_{i=1}^3 \mathbb{E}\lambda\{B_t : t \in [i-1, i]\} = 3\mathbb{E}X .$$

From the above, assuming $\mathbb{E}X < \infty$, we deduce that

$$\mathbb{E} \left[\lambda\{B_t : t \in [0, 3]\} - \sum_{i=1}^3 \lambda\{B_t : t \in [i-1, i]\} \right] = 0 ,$$

and since this variable is nonnegative,

$$\lambda\{B_t : t \in [0, 3]\} = \sum_{i=1}^3 \lambda\{B_t : t \in [i-1, i]\} = 0 \text{ almost surely .}$$

In particular,

$$\mathbb{E}\lambda\{\{B_t : t \in [0, 1]\} \cap \{B_t : t \in [2, 3]\}\} = 0 . \quad (27)$$

To justify this, we need to show that X has finite mean. If $X > a$ then Brownian motion has to leave a box of side-length $\sqrt{a}$ centered at the origin. Therefore at least one of the two coordinates of B_t must exceed $\sqrt{a}$ in absolute value for $t \in [0, 1]$. Setting $M_t^{(i)} = B_t^{(i)}$ for $i = 1, 2$,

$$\mathbb{P}(X > a) \leq \mathbb{P}\left(|M_1^{(i)}| > \sqrt{a} \text{ for some } i = 1, 2\right) \leq 4\mathbb{P}\left(M_1^{(1)} > \sqrt{a}\right) .$$

By the reflection principle, we obtain

$$\mathbb{P}(X > a) \leq 8\mathbb{P}\left(B_1^{(1)} > \sqrt{a}\right) .$$

Therefore

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > a) \, da \leq 8 \int_0^\infty \mathbb{P}\left(B_1^{(1)} > \sqrt{a}\right) \, da$$

and since $B_1^{(1)}$ is a standard Gaussian, this is finite.

Now we define two Brownian motions:

$$W_1(t) = B_t \text{ and } W_2(t) = (B_{t+2} - B_2) + B_1 \text{ for } t \in [0, 1] .$$

By independence of increments, both of these are independent of $Y := B_2 - B_1$ (the second is a sum of increments B_1 and $B_{t+2} - B_2$). Letting

$$R(x) = \lambda\{\{W_1(t) : t \in [0, 1]\} \cap x + \{W_2(t) : t \in [0, 1]\}\} ,$$

we have

$$R(Y) = \lambda \{ \{B_t : t \in [0, 1]\} \cap \{B_{t+2} : t \in [0, 1]\} \} .$$

By (27),

$$0 = \mathbb{E}R(Y) = \mathbb{E} [\mathbb{E}[R(Y) \mid Y]] = \frac{1}{\sqrt{2\pi}} \int e^{-x^2/2} \mathbb{E}R(x) \, dx ,$$

So $\mathbb{E}R(x) = 0$ for almost all $x \in \mathbb{R}^2$ and therefore $R(x) = 0$ for almost all $x \in \mathbb{R}^2$, almost surely. This means

$$\lambda\{x \in \mathbb{R}^2 : R(x) > 0\} = 0 \text{ almost surely} . \quad (28)$$

From (28), we will use a general lemma to finish the proof. The bounded assumption below is not necessary, but it is ok for us since the sets A and B are just images of Brownian motion.

Lemma 14.4. *If $A, B \subset \mathbb{R}^2$ are bounded Borel sets with positive area,*

$$\lambda \{ x \in \mathbb{R}^2 : \lambda\{A \cap (x + B)\} > 0 \} > 0 .$$

Proof. By Fubini,

$$\begin{aligned} \int (\mathbf{1}_A * \mathbf{1}_{-B})(x) \, dx &= \int \int \mathbf{1}_A(w) \mathbf{1}_{-B}(x - w) \, dw \, dx \\ &= \int \int \mathbf{1}_A(w) \mathbf{1}_B(w - x) \, dw \, dx \\ &= \int \mathbf{1}_A(w) \left(\int \mathbf{1}_B(w - x) \, dx \right) \, dw \\ &= \lambda(A) \lambda(B) , \end{aligned}$$

and this is positive. Therefore

$$\lambda\{x \in \mathbb{R}^2 : (\mathbf{1}_A * \mathbf{1}_{-B})(x) > 0\} > 0 .$$

However

$$(\mathbf{1}_A * \mathbf{1}_{-B})(x) = \int \mathbf{1}_A(w) \mathbf{1}_{-B}(x - w) \, dw = \int \mathbf{1}_A(w) \mathbf{1}_{B+x}(w) \, dw = \lambda\{A \cap (x + B)\} .$$

□

By the lemma applied to (28), we see that almost surely, either $\{W_1(t) : t \in [0, 1]\}$ or $\{W_2(t) : t \in [0, 1]\}$ has Lebesgue measure zero. Note

$$\lambda\{W_2(t) : t \in [0, 1]\} = \lambda\{B_{2+t} - B_2 + B_1 : t \in [0, 1]\} = \lambda\{B_{2+t} - B_t : t \in [0, 1]\}$$

and this is independent of

$$\lambda\{W_1(t) : t \in [0, 1]\} = \lambda\{B_t : t \in [0, 1]\}$$

with the same distribution. Therefore they must both be zero with probability one and this completes the proof. □

LECTURES 30 AND 31

From the result on area, we can prove that for $d \geq 2$, Brownian motion is not point recurrent.

Proof. First take $d = 2$. By the previous theorem, for fixed y ,

$$\int \mathbb{P}_y(x \in \{B_t : t \in [0, 1]\}) \, dx = 0 .$$

This means that for λ -almost every $x \in \mathbb{R}^2$,

$$\begin{aligned} 0 &= \mathbb{P}_y(x \in \{B_t : t \in [0, 1]\}) = \mathbb{P}_0(x - y \in \{B_t : t \in [0, 1]\}) \\ &= \mathbb{P}_0(y - x \in \{B_t : t \in [0, 1]\}) = \mathbb{P}_x(y \in \{B_t : t \in [0, 1]\}) . \end{aligned}$$

Therefore for $n \geq 1$, setting $B_t^* = B_{t+1/n} - B_{1/n}$,

$$\begin{aligned} \mathbb{P}_0(y \in \{B_t : t \in [1/n, 1]\}) &= \mathbb{E}_0 [\mathbb{P}_0(y \in \{B_t : t \in [1/n, 1]\}) \mid B_{1/n}] \\ &= \mathbb{E}_0 [\mathbb{P}_0(y \in (B_{1/n} + \{B_t^* : t \in [0, 1 - 1/n]\})) \mid B_{1/n}] \\ &= \mathbb{E}_0 [\mathbb{P}_{B_{1/n}}(y \in \{B_t : t \in [0, 1 - 1/n]\})] \\ &= \int \mathbb{P}_x(y \in \{B_t : t \in [0, 1 - 1/n]\}) \, dB_{1/n}(x) = 0 . \end{aligned}$$

To go from the second to third line we have used the fact that B_t and $(B_t^* : t \in [0, 1 - 1/n])$, so conditioning on B_t amounts to viewing B_t as a deterministic number. We finish with

$$\mathbb{P}_0(y \in \{B_t : t \in [0, 1]\}) = \lim_n \mathbb{P}_0(y \in \{B_t : t \in [1/n, 1]\}) = 0 .$$

If $d > 2$ then just project Brownian motion onto the first two coordinates. $\square$

- Although in $d = 2$ Brownian motion is not point recurrent, it is neighborhood recurrent: for any ball $\mathcal{B}_r$ of radius $r > 0$, and all $y \in \mathbb{R}^2$,

$$\mathbb{P}_y(B_t \in \mathcal{B}_r \text{ for some } t \geq 0) = 1 .$$

- For $d \geq 3$, Brownian motion is neither point recurrent nor neighborhood recurrent. These statements can be proved using the relation to harmonic functions. See Peres-Mörters, Ch. 3.

15 Stochastic integration

One way to build continuous stochastic processes from Brownian motion is to integrate. For example, since Brownian paths are continuous, we can always integrate them path by path to obtain a new process

$$X_t = \int_0^t B_s \, ds .$$

We can give many variations on this, like integrating functions of Brownian motion: $\int_0^t f(B_s) ds$. Both of these examples have B_t as the integrand. It is much more difficult to make sense of expressions like $\int_0^t f(s) dB_s$ or $\int_0^t X_s dB_s$, where X_s is another stochastic process. Here we are integrating with respect to Brownian motion.

Why would we want to define such integrals? Well one point of view comes from stochastic differential equations. Suppose X_t is a random process which, say, models the erratic movement of a particle. Suppose we can write an equation like

$$X_t = \int_0^t a(X_s) dB_s + \int_0^t b(X_s) ds .$$

This is written in short hand (as an SDE) as

$$dX_t = a(X_t) dB_t + b(X_t) dt .$$

We can interpret this infinitesimally by thinking that in a small amount of time, our particle gets a push in the direction of $b(X_s)$ (where X_s is the current position) – this is the term on the right, which is the *drift term* – but also a random diffusive push by a Brownian motion B_t with strength $a(X_s)$ – this is the *diffusive term*. Here we are modeling a system that satisfies an ODE that is subject to some random noise.

The problem here of course is that dB_s does not make any sense. We know from the theory of Lebesgue-Stieltjes integrals that we can only define $\int_0^t f(s) dg(s)$ when g has bounded variation. And we have seen that Brownian motion does not have this property: functions of bounded variation are differentiable almost everywhere, and Brownian motion is not differentiable anywhere. So to make sense of this integral we will have to take advantage of the fact that, since we are dealing with randomness, we can build integrals using various types of limits, including almost sure convergence, probability convergence, convergence in L^p , etc.

15.1 Integration of step processes

As usual, we will give a simple definition of the integral when the integrand is itself simple. For the remainder, let (B_t) be a Brownian motion defined on a space $(\Omega, \Sigma, \mathbb{P})$ and $(\mathcal{F}_t)$ any filtration such that

1. (B_t) is adapted to $(\mathcal{F}_t)$,
2. the strong Markov property holds for (B_t) on $(\mathcal{F}_t)$ and
3. $\sigma(\mathcal{F}_t : t \geq 0)$ contains all $\mathbb{P}$ -probability zero events in Σ (it is *complete*).

Definition 15.1. *A simple process is of the form*

$$S_t = S_t(\omega) = \sum_{j=1}^k a_j(\omega) \mathbf{1}_{(t_j, t_{j+1}]}(t) ,$$

where a_j is an $\mathcal{F}_{t_j}$ -measurable random variable and $0 \leq t_1 \leq \dots \leq t_{k+1}$. For a simple process we define the stochastic integral

$$\int_0^\infty S_s dB_s = \sum_{j=1}^k a_j (B_{t_{j+1}} - B_{t_j}) .$$

We should check that the integral of a simple process is well-defined. So suppose that (S_t) can be written in two ways:

$$S_t = \sum_{j=1}^k a_j \mathbf{1}_{(t_j, t_{j+1}]} \text{ and } S_t = \sum_{i=1}^m b_i \mathbf{1}_{(u_i, u_{i+1}]} .$$

We must show that

$$\sum_{j=1}^k a_j (B_{t_{j+1}} - B_{t_j}) = \sum_{i=1}^m b_i (B_{u_{i+1}} - B_{u_i}) . \quad (29)$$

Write $A = \{t_1, \dots, t_{k+1}\}$ and $B = \{u_1, \dots, u_{m+1}\}$. First assume that $A \subset B$. If $\#B = \#A + 1$ then there is some interval, say $(t_1, t_2]$, with $u_1 = t_1, u_3 = t_2$ and $u_2 \in (t_1, t_2]$. It follows that $b_2 = a_1$ and so (29) holds. If $\#B \geq \#A + 1$ then we simply iterate this argument: we write $B \setminus A = \{u_{k_1}, \dots, u_{k_s}\}$ and apply the previous argument to move from A to $A \cup \{u_{k_1}\}$ to $A \cup \{u_{k_1}, u_{k_2}\}$ and so on.

For general A and B , we let $C = A \cup B = \{r_1, \dots, r_{p+1}\}$ and write

$$S_t = \sum_{s=1}^p S_{r_s} \mathbf{1}_{(r_s, r_{s+1}]} .$$

Then by the above argument applied to A and C , then to B and C , we obtain

$$\sum_{j=1}^k a_j (B_{t_{j+1}} - B_{t_j}) = \sum_{s=1}^p S_{r_s} (B_{r_{s+1}} - B_{r_s}) = \sum_{i=1}^m b_i (B_{u_{i+1}} - B_{u_i}) .$$

Note the following facts.

- A simple process is adapted. That is, if (S_t) is simple, then S_t is $\mathcal{F}_t$ -measurable.
- A simple process is jointly measurable. In fact, it is *progressively measurable*. That is, for each $t \geq 0$, the map $S : [0, t] \times \Omega \rightarrow \mathbb{R}$ given by $S(s, \omega) = S_s(\omega)$ is measurable relative to $\mathcal{B}([0, t]) \times \mathcal{F}_t$, where $\mathcal{B}([0, t])$ is the Borel sigma-algebra of $[0, t]$.
- Actually all processes that are adapted to $(\mathcal{F}_t)$ and are either right or left continuous are progressively measurable.

Proof. The proof is very much like the one that Brownian motion is jointly measurable. If X_t is right-continuous and adapted to $(\mathcal{F}_t)$, then for $0 \leq s \leq t$ define $X_s^{(n)}$ by

$$X_s^{(n)} = \begin{cases} X_0 & \text{if } s = 0 \\ X_{\frac{(k+1)t}{2^n}} & \text{if } \frac{kt}{2^n} < s \leq \frac{(k+1)t}{2^n} \end{cases} .$$

As before we can show that this function is $\mathcal{B}([0, t]) \times \mathcal{F}_t$ measurable. By right continuity, $X_s^{(n)} \rightarrow X_s$ for all $s \in [0, t]$ and $\omega \in \Omega$, so the limit is $\mathcal{B}([0, t]) \times \mathcal{F}_t$ -measurable. The proof for left continuity is similar. $\square$

- If (X_t) is progressively measurable then for each t , the $t \mapsto \mathbb{E}(X_t)_+$ and $t \mapsto \mathbb{E}(X_t)_-$ are Lebesgue measurable on all intervals. Therefore by Fubini,

$$\mathbb{E} \int_a^b f(X_t) dt = \int_a^b \mathbb{E} f(X_t) dt \text{ if } \int_a^b \mathbb{E} |f(X_t)| dt < \infty .$$

Let us give some important properties of this simple definition. We will restrict to simple processes that have square integrable jumps. That is, we define

$$\|H\|_2^2 = \int_0^\infty \mathbb{E} H_t^2 dt .$$

Theorem 15.2. *Let S_t, T_t be simple processes such that $\|S\|_2, \|T\|_2 < \infty$.*

1. *For $a, b \in \mathbb{R}$,*

$$\int_0^\infty (aS_t + bT_t) dB_t = a \int_0^\infty S_t dB_t + b \int_0^\infty T_t dB_t .$$

2. *The mean $\mathbb{E} \int_0^\infty S_t dB_t$ is zero.*

3. *(Isometry)*

$$\mathbb{E} \left(\int_0^\infty S_t dB_t \right)^2 = \mathbb{E} \int_0^\infty S_t^2 dB_t .$$

Proof. If S_t and T_t are simple processes defined over the same intervals:

$$S_t = \sum_{j=1}^k a_j \mathbf{1}_{(t_j, t_{j+1}]}(t) \text{ and } T_t = \sum_{j=1}^k b_j \mathbf{1}_{(t_j, t_{j+1}]}(t)$$

then $S_t + T_t = \sum_{j=1}^k (a_j + b_j) \mathbf{1}_{(t_j, t_{j+1}]}(t)$ and

$$\int_0^\infty (S_t + T_t) dB_t = \sum_{j=1}^k (a_j + b_j) (B_{t_{j+1}} - B_{t_j}) = \int_0^\infty S_t dB_t + \int_0^\infty T_t dB_t .$$

If S_t and T_t are not defined over the same intervals, we can force this. Namely, if $S_t = \sum_{j=1}^k a_j \mathbf{1}_{(t_j, t_{j+1}]}$ and $T_t = \sum_{j=1}^m b_j \mathbf{1}_{(s_j, s_{j+1}]}$ then we can write $A = \{t_1, \dots, t_{k+1}\}$, $B = \{s_1, \dots, s_{m+1}\}$ and $C = A \cup B = \{r_1, \dots, r_{p+1}\}$. Then we can write our simple processes as

$$S_t = \sum_{i=1}^p S_{r_i} \mathbf{1}_{(r_i, r_{i+1}]} \text{ and } T_t = \sum_{i=1}^p T_{r_i} \mathbf{1}_{(r_i, r_{i+1}]}$$

and this puts us back in the previous situation, proving additivity.

Also if $S_t = \sum_{j=1}^k a_j \mathbf{1}_{(t_j, t_{j+1}]}$ and $a \in \mathbb{R}$ then

$$\int_0^\infty a S_t \, dB_t = \sum_{j=1}^k a a_j (B_{t_{j+1}} - B_{t_j}) = a \int_0^\infty S_t \, dB_t .$$

This shows part 1.

For part 2, note that $B_{t_{j+1}} - B_{t_j}$ is independent of $\mathcal{F}_j$ and therefore of a_j , so

$$\mathbb{E} \int_0^\infty S_t \, dB_t = \sum_{j=1}^k \mathbb{E} a_j \mathbb{E} (B_{t_{j+1}} - B_{t_j}) = 0 .$$

Last if $y_j = a_j (B_{t_{j+1}} - B_{t_j})$, then by independence, $\mathbb{E} y_j^2 = \mathbb{E} a_j^2 \mathbb{E} (B_{t_{j+1}} - B_{t_j})$, which is $\mathbb{E} a_j^2 (t_{j+1} - t_j)$. Therefore

$$\sum_{j=1}^k \mathbb{E} y_j^2 = \sum_{j=1}^k \mathbb{E} a_j^2 (t_{j+1} - t_j) = \int_0^\infty \mathbb{E} S_t^2 \, dt .$$

Now if $i < j$ then

$$\mathbb{E} y_i y_j = \mathbb{E} [\mathbb{E} [y_i y_j \mid \mathcal{F}_{t_j}]] = \mathbb{E} [y_i a_j \mathbb{E} [B_{t_{j+1}} - B_{t_j} \mid \mathcal{F}_{t_j}]] = 0 .$$

Therefore

$$\mathbb{E} \left(\int_0^\infty S_t \, dB_t \right)^2 = \sum_{i,j=1}^k \mathbb{E} y_i y_j = \int_0^\infty \mathbb{E} S_t^2 \, dt .$$

□

LECTURE 32

15.2 Approximation of progressively measurable integrands

We can use the isometry to define integrals for more processes.

Definition 15.3. Define $\mathcal{V}$ as the collection of progressively measurable processes X_t with $\|X\|_2 < \infty$ such that there exists a sequence $(S_t^{(n)})$ of simple processes with $\|S^{(n)}\|_2 < \infty$ for all n and

$$\|S^{(n)} - X\|_2 \rightarrow 0 . \tag{30}$$

Write $\mathcal{M}$ for the collection of progressively measurable processes (X_t) with $\|X\|_2 < \infty$.

You will show on the homework that $\mathcal{V} = \mathcal{M}$. The idea is first to approximate by a bounded progressively measurable process, then a continuous one, and then a simple one.

Lemma 15.4.

$$\mathcal{V} = \mathcal{M} .$$

To show existence of the integral, we will continuously extend the definition from simple processes to the class $\mathcal{M}$. This is analogous to the following situation. Suppose that X and Y are metric spaces and $T : \hat{X} \rightarrow Y$ is an isometry defined on a dense subset $\hat{X}$ of X . Then one can extend T to all of X simply by defining $T(x) = \lim_n T(x_n)$ for any sequence (x_n) in $\hat{X}$ such that $x_n \rightarrow x$. By the isometry property of T , this definition does not depend on the sequence (x_n) chosen.

To realize the above strategy, we would like T to be the map from $\mathcal{M}$ to L^2 that sends a process to its integral. We do not need to fully show that $\mathcal{M}$ is a metric space; we only need the following fact: if $X, Y \in \mathcal{M}$ then the triangle inequality holds:

$$\|X + Y\|_2 = \sqrt{\int_0^\infty \mathbb{E}(X_t + Y_t)^2 dt} \leq \sqrt{\int_0^\infty \mathbb{E}X_t^2 dt} + \sqrt{\int_0^\infty \mathbb{E}Y_t^2 dt} = \|X\|_2 + \|Y\|_2 .$$

To see why the middle inequality holds, just note that it is the triangle inequality using the L^2 -norm on a bigger space: $\Omega \times [0, \infty)$ with measure $\mathbb{P} \times \lambda$, where λ is Lebesgue measure on $[0, \infty)$.

Theorem 15.5 (Existence of the integral). *Let $(X_t) \in \mathcal{M}$. The limit*

$$\int_0^\infty X_t dB_t := \lim_n \int_0^\infty S_t^{(n)} dB_t \text{ exists}$$

in $L^2(\Omega)$ for any sequence $(S_t^{(n)})$ of simple processes as in (30) and is independent of the sequence. Furthermore,

$$\mathbb{E} \left(\int_0^\infty X_t dB_t \right)^2 = \int_0^\infty \mathbb{E}X_t^2 dt . \quad (31)$$

Proof. If $(S_t^{(n)})$ and $(T_t^{(n)})$ are sequences of simple processes as in (30), then for any $m, n \geq 1$, $(S_t^{(m)} - S_t^{(n)})$ is simple, so by isometry,

$$\left\| \int_0^\infty S_t^{(n)} dB_t - \int_0^\infty S_t^{(m)} dB_t \right\|_2 = \|S^{(n)} - S^{(m)}\|_2 \leq \|S^{(n)} - X\|_2 + \|S^{(m)} - X\|_2 \rightarrow 0 .$$

Note that on the left side, we are using the L^2 norm on the space $L^2(\Omega, \mathbb{P})$, since the integrals are just random variables. This shows the sequence $\int_0^\infty S_t^{(n)} dB_t$ is Cauchy in L^2 and, by completeness, it has a limit. Also

$$\left\| \int_0^\infty S_t^{(n)} dB_t - \int_0^\infty T_t^{(n)} dB_t \right\|_2 = \|S^{(n)} - T^{(n)}\|_2 \leq \|S^{(n)} - X\|_2 + \|T^{(n)} - X\|_2 \rightarrow 0 ,$$

so this limit does not depend on the sequence of simple processes used.

To show isometry, note by definition of the integral,

$$\left\| \int_0^\infty S_t^{(n)} dB_t - \int_0^\infty X_t dB_t \right\|_2 \rightarrow 0 .$$

By the triangle inequality and isometry,

$$\begin{aligned} \left| \|S^{(n)}\|_2 - \left\| \int_0^\infty X_t dB_t \right\|_2 \right| &= \left| \left\| \int_0^\infty S_t^{(n)} dB_t \right\|_2 - \left\| \int_0^\infty X_t dB_t \right\|_2 \right| \\ &\leq \left\| \int_0^\infty (S_t^{(n)} - X_t) dB_t \right\|_2 \rightarrow 0 . \end{aligned}$$

Thus $\|S^{(n)}\|_2 \rightarrow \left\| \int_0^\infty X_t dB_t \right\|_2$. However by the triangle inequality proved before this theorem,

$$|\|S^{(n)}\|_2 - \|X\|_2| \leq \|S^{(n)} - X\|_2 \rightarrow 0 ,$$

so also $\|S^{(n)}\|_2 \rightarrow \|X\|_2$. This shows (31). $\square$

15.3 Integral as a process

For $T > 0$, let $\mathcal{M}_T$ be the set of progressively measurable $(X_t)_{t \geq 0}$ such that

$$\int_0^T \mathbb{E} X_s^2 ds < \infty .$$

For such (X_t) , we can define the integral of X_t over $t \in [0, T]$ as a process. This is analogous to defining an indefinite integral.

Definition 15.6. *The indefinite integral of $(X_t) \in \mathcal{M}_T$ is defined to be the process $(Y_t)_{t \in [0, T]}$ given by*

$$Y_t = \int_0^t X_s dB_s := \int_0^\infty X_s \mathbf{1}_{[0, t]}(s) dB_s .$$

Note that for each $t \in [0, T]$, the process $(X_s \mathbf{1}_{[0, t]}(s)) \in \mathcal{M}_t$ since it is a product of two progressively measurable processes ($\mathbf{1}_{[0, t]}(s)$ is because it is deterministic) and by assumption

$$\int_0^\infty \mathbb{E} (X_s \mathbf{1}_{[0, t]}(s))^2 ds = \int_0^t \mathbb{E} X_s^2 ds \leq \int_0^T \mathbb{E} X_s^2 ds < \infty .$$

So the integral is defined for all $t \in [0, T]$.

Just as in the case of constructing Brownian motion, we must modify (Y_t) to get a version with continuous paths.

Theorem 15.7. *For $(X_t) \in \mathcal{M}_T$ and $T > 0$, the process $(Y_t)_{t \in [0, T]}$ given by $Y_t = \int_0^t X_s dB_s$ is adapted to $(\mathcal{F}_t)$ and has a continuous modification. That is, there is a process $(J_t)_{t \in [0, T]}$ with continuous paths such that*

$$\mathbb{P}(J_t = Y_t) = 1 \text{ for all } t \geq 0 .$$

Furthermore, (J_t) is a martingale.

LECTURE 33

Now we will prove the theorem

Theorem 15.8. *For $(X_t) \in \mathcal{M}_T$ and $T > 0$, the process $(Y_t)_{t \in [0, T]}$ given by $Y_t = \int_0^t X_s dB_s$ is adapted to $(\mathcal{F}_t)$ and has a continuous modification. That is, there is a process $(J_t)_{t \in [0, T]}$ with continuous paths such that*

$$\mathbb{P}(J_t = Y_t) = 1 \text{ for all } t \geq 0 .$$

Furthermore, (J_t) is a martingale.

Proof. To begin with, let us show (Y_t) is adapted to $\mathcal{F}_t$. To see this, let $(S_s^{(n)})_{s \geq 0}$ be a sequence of simple processes such that $\|X \mathbf{1}_{[0, t]} - S^{(n)}\|_2 \rightarrow 0$. Although $S^{(n)}$ is defined using intervals $[t_j, t_{j+1})$ for $0 = t_0 < t_1 < \dots < t_{k+1}$, we can “chop” it at time t , defining a new step process $T^{(n)}$ by $T^{(n)} = S^{(n)} \mathbf{1}_{[0, t]}$. Then by isometry,

$$\begin{aligned} \mathbb{E} \left(\int_0^t T_s^{(n)} dB_s - \int_0^t X_s dB_s \right)^2 &= \int_0^\infty \mathbb{E} (T_s^{(n)} - X_s \mathbf{1}_{[0, t]}(s))^2 ds \\ &\leq \|S^{(n)} - X\|_2^2 \rightarrow 0 \end{aligned}$$

However the term on the left is $\mathbb{E} \left(\int_0^t T_s^{(n)} dB_s - Y_t \right)^2$, so

$$\int_0^t T_s^{(n)} dB_s \rightarrow Y_t \text{ in } L^2 .$$

Since $\int_0^t T_s^{(n)} dB_s$ is $\mathcal{F}_t$ measurable, so is Y_t .

Next we move toward the continuous modification. We will first approximate by simple processes and show their integrals are martingales with continuous paths. Then we will apply Doob’s maximal inequality to exhibit (Y_t) as a uniform limit of continuous functions.

So pick a sequence of simple processes $(S_t^{(n)}) \in \mathcal{M}$ such that $\|S^{(n)} - X \mathbf{1}_{[0, T]}\|_2 \rightarrow 0$:

$$\int_0^T \mathbb{E} (S_t^{(n)} - X_t)^2 dt \rightarrow 0 .$$

Now define

$$I_t^{(n)} = \int_0^t S_s^{(n)} dB_s \text{ for } t \in [0, T] .$$

We claim that for each n , $(I_t^{(n)})$ is a martingale relative to $(\mathcal{F}_t)$. To prove this note first that $(I_t^{(n)})$ is adapted to $(\mathcal{F}_t)$, and write

$$S_t^{(n)} = \sum_{j=1}^k a_j \mathbf{1}_{(t_j, t_{j+1}]}(t)$$

and let s, t satisfy $0 \leq s \leq t \leq T$. If $s \in (t_p, t_{p+1}]$ and $t \in (t_q, t_{q+1}]$ for $p \leq q$, then

$$\begin{aligned} I_t^{(n)} - I_s^{(n)} &= \int_0^\infty S_u^{(n)} \mathbf{1}_{[0,t]}(u) \, dB_u - \int_0^\infty S_u^{(n)} \mathbf{1}_{[0,s]}(u) \, dB_u \\ &= \int_0^\infty S_u^{(n)} \mathbf{1}_{(s,t]}(u) \, dB_u \\ &= a_p(B_{t_{p+1}} - B_s) + a_{p+1}(B_{t_{p+2}} - B_{t_{p+1}}) + \cdots + a_q(B_t - B_{t_q}) . \end{aligned}$$

For the first term,

$$\mathbb{E}[a_p(B_{t_{p+1}} - B_s) \mid \mathcal{F}_s] = a_p \mathbb{E}(B_{t_{p+1}} - B_s) = 0 .$$

All subsequent terms are of the form $A(B_{u_2} - B_{u_1})$ for $s \leq u_1 \leq u_2$ and A a random variable which is $\mathcal{F}_{u_1}$ -measurable. So using the tower property,

$$\begin{aligned} \mathbb{E}[A(B_{u_2} - B_{u_1}) \mid \mathcal{F}_s] &= \mathbb{E}[\mathbb{E}[A(B_{u_2} - B_{u_1}) \mid \mathcal{F}_{u_1}] \mid \mathcal{F}_s] \\ &= \mathbb{E}[A \mathbb{E}[B_{u_2} - B_{u_1}]] = 0 . \end{aligned}$$

Therefore

$$\mathbb{E} \left[\int_0^t S_u^{(n)} \, dB_u \mid \mathcal{F}_s \right] = \int_0^s S_u^{(n)} \, dB_u + \mathbb{E} \left[\int_0^t S_u^{(n)} \, dB_u - \int_0^s S_u^{(n)} \, dB_u \mid \mathcal{F}_s \right] .$$

and this is $\int_0^s S_u^{(n)} \, dB_u$, proving the martingale property.

Furthermore, $t \mapsto \int_0^t S_s^{(n)} \, dB_s$ is continuous by the representation above for $\int_0^t S_u^{(n)} \, dB_u - \int_0^s S_u^{(n)} \, dB_u$. So we will analyze it using the continuous time version of the Doob maximal inequality.

Lemma 15.9. *Let $(Y_t)_{t \in [0, T]}$ be a martingale relative to $(\mathcal{F}_t)$ such that $t \mapsto Y_t$ is continuous. Then for each $a \geq 0$, $t \in [0, T]$,*

$$\mathbb{P} \left(\max_{0 \leq s \leq t} |Y_s| \geq a \right) \leq \frac{1}{a^2} \mathbb{E} Y_t^2 .$$

Proof. Define for each n the sequence $(X_0^{(n)}, \mathcal{F}_0^{(n)}), \dots, (X_{2^n}^{(n)}, \mathcal{F}_{2^n}^{(n)})$ by

$$X_k^{(n)} = Y_{\frac{kt}{2^n}} \text{ and } \mathcal{F}_k^{(n)} = \mathcal{F}_{\frac{kt}{2^n}} .$$

Then for each n , this sequence is a martingale. So by the maximal inequality for martingale sequences,

$$\mathbb{P} \left(\max_{k=0, \dots, 2^n} |X_k^{(n)}| \geq a \right) \leq \frac{1}{a^2} \mathbb{E} |X_{2^n}^{(n)}| = \frac{1}{a^2} \mathbb{E} Y_t^2 .$$

As n increases, the events on the left are nested. So if we take $n \rightarrow \infty$, we obtain for $D_t = \{kt/2^n : n \geq 1, k = 0, \dots, 2^n\}$

$$\mathbb{P} \left(\sup_{q \in D_t \cap [0, t]} |Y_q| \geq a \right) \leq \frac{1}{a^2} \mathbb{E} Y_t^2 .$$

Since Y_t is continuous, the left side equals $\mathbb{P}(\max_{0 \leq s \leq t} |Y_s| \geq a)$ so we are done. $\square$

Using Doob's inequality, for $m, n \geq 1$, since $(I_t^{(n)} - I_t^{(m)})_{t \in [0, T]}$ is a continuous martingale, if $a > 0$,

$$\mathbb{P} \left(\max_{0 \leq t \leq T} |I_s^{(n)} - I_s^{(m)}| > a \right) \leq \frac{1}{a^2} \mathbb{E}(I_T^{(n)} - I_T^{(m)})^2 = \frac{1}{\epsilon^2} \int_0^T \mathbb{E}(S_t^{(n)} - S_t^{(m)})^2 dt .$$

Because this converges to 0 as $n, m \rightarrow \infty$ (by choice of the approximating sequence $(S^{(n)})$), we may choose a subsequence (n_k) such that

$$\mathbb{P} \left(\max_{0 \leq t \leq T} |I_t^{(n_k)} - I_t^{(n_{k+1})}| > 2^{-k} \right) \leq 2^{-k}$$

and by Borel-Cantelli

$$\mathbb{P} \left(\max_{0 \leq t \leq T} |I_t^{(n_k)} - I_t^{(n_{k+1})}| > 2^{-k} \text{ i.o.} \right) = 0 .$$

Therefore for almost every ω , there is a $K(\omega)$ such that

$$\max_{0 \leq t \leq T} |I_t^{(n_{k+1})} - I_t^{(n_k)}| \leq 2^{-k} \text{ for } k \geq K(\omega) .$$

This means that the sequence $(I_t^{(n_k)})_k$ of functions converges uniformly, and so the limit, called $(J_t)_{t \in [0, T]}$ is continuous.

Next time we will show that (J_t) is a modification of (Y_t) and that it is a martingale. $\square$

LECTURE 34

First we finish proving:

Theorem 15.10. *For $(X_t) \in \mathcal{M}_T$ and $T > 0$, the process $(Y_t)_{t \in [0, T]}$ given by $Y_t = \int_0^t X_s dB_s$ is adapted to $(\mathcal{F}_t)$ and has a continuous modification. That is, there is a process $(J_t)_{t \in [0, T]}$ with continuous paths such that*

$$\mathbb{P}(J_t = Y_t) = 1 \text{ for all } t \geq 0 .$$

Furthermore, (J_t) is a martingale.

Proof. We have shown everything but that

$$J_t := \lim_k I_t^{(n_k)}$$

(where $I_t^{(n)}$ is an indefinite integral of a simple process $S^{(n)}$ that approximates X , and n_k is a subsequence chosen so that the convergence above is uniform almost surely) is a modification of Y_t and that it is a martingale.

To see that for each $t \in [0, T]$, $J_t = Y_t$ almost surely, we note that for fixed t , $I_t^{(n)} \rightarrow Y_t$ in L^2 : by isometry,

$$\|I_t^{(n)} - Y_t\|_2 = \|(S^{(n)} - X)\mathbf{1}_{[0,t]}\|_2 \leq \|(S^{(n)} - X)\mathbf{1}_{[0,T]}\|_2 \rightarrow 0 .$$

and so $I_t^{(n_k)} \rightarrow Y_t$ in L^2 as well. Thus there is a subsequence $I_t^{(n_{k_j})}$ that converges to Y_t almost surely. However $I_t^{(n_k)}$ converges almost surely to J_t by the above argument, so we must have $J_t = Y_t$ almost surely.

Last we must verify that (J_t) is a martingale. We already showed that for each n , $(I_t^{(n)})_{t \in [0, T]}$ is a martingale, so for $0 \leq s \leq t \leq T$,

$$\mathbb{E} \left[I_t^{(n)} \mid \mathcal{F}_s \right] = I_s^{(n)} \text{ almost surely .}$$

Because $I_t^{(n)} \rightarrow J_t$ in L^2 , using Jensen,

$$\mathbb{E} \left[\left(\mathbb{E}[I_t^{(n)} \mid \mathcal{F}_s] - \mathbb{E}[J_t \mid \mathcal{F}_s] \right)^2 \right] \leq \mathbb{E}(I_t^{(n)} - J_t)^2 \rightarrow 0 .$$

This gives

$$\mathbb{E}[I_t^{(n)} \mid \mathcal{F}_s] \rightarrow \mathbb{E}[J_t \mid \mathcal{F}_s] \text{ in } L^2 .$$

But the left side is $I_s^{(n)}$, which converges in L^2 to J_s , so

$$\mathbb{E}[J_t \mid \mathcal{F}_s] = J_s \text{ almost surely .}$$

□

16 Ito's formula

Because of the last section, if we are given a process X_t that is in $\mathcal{M}_T$ for some $T > 0$ and, say, a continuous function $f : \mathbb{R} \rightarrow \mathbb{R}$, we can make sense of

$$Y_t := \int_0^t X_s \, dB_s + \int_0^t f(B_t) \, dt .$$

In this case we would write this in short-hand as

$$dY_t = X_t dB_t + f(B_t) dt .$$

Now we ask if we can write a similar short-hand for functions of Y_t . That is, if $f : \mathbb{R} \rightarrow \mathbb{R}$ then what is $df(Y_t)$? We will start with the case that $Y_t = B_t$. So we ask if we can write

$$df = X_t dB_t + Z_t dt .$$

In integral form, this is

$$f(B_t) - f(B_0) = \int_0^t X_s \, dB_s + \int_0^t Z_s \, ds .$$

To relate to ordinary calculus, if $B_t = g(t)$ were a deterministic differentiable function, we could use the chain rule for

$$(f \circ g)'(t) = f'(g(t)) \cdot g'(t)$$

so

$$f(g(t)) - f(g(0)) = \int_0^t f'(g) \, dg .$$

In the case that we use Brownian motion for g , the non-differentiability leads to an extra term.

Theorem 16.1 (Ito's formula – simplified version). *Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be C^1 such that for some $T > 0$, $\int_0^T \mathbb{E}(f'(B_t))^2 \, dt < \infty$. Then with probability one,*

$$f(B_t) - f(B_0) = \int_0^t f'(B_s) \, dB_s + \frac{1}{2} \int_0^t f''(B_s) \, ds \text{ for all } t \in [0, T] . \quad (32)$$

Before we give the proof, let's give some examples and discussion. This version of Ito's formula can be written in differential form as

$$df = f' dB + \frac{1}{2} f'' dt .$$

You can remember the proof as follows. We do a Taylor expansion for f :

$$f(B_{t+h}) - f(B_t) = f'(B_t)(B_{t+h} - B_t) + \frac{1}{2} f''(B_t)(B_{t+h} - B_t)^2 + \dots .$$

This is written in differential notation as

$$df = f' \, dB + \frac{1}{2} f'' \, (dB)^2 + \dots .$$

Now we use the rule $(dB)^2 = dt$ and $(dB)^k = 0$ for $k \geq 3$. This is justified because we know that

$$\sum_{k=1}^{k_n} (B_{t_k^{(n)}} - B_{t_{k-1}^{(n)}})^2 \rightarrow t \text{ and } \sum_{k=1}^{k_n} (B_{t_k^{(n)}} - B_{t_{k-1}^{(n)}})^m \rightarrow 0 \text{ for } m \geq 3 .$$

Therefore we obtain

$$df = f' dB + \frac{1}{2} f'' dt .$$

Using this heuristic we will “derive” Ito's formula for functions of Brownian motion and time. The actual proof is very similar to the previous one, and we will not use it later, so we will just show it for some examples.

Theorem 16.2 (Ito's formula). *Let $f : [0, \infty) \times \mathbb{R}$ be twice continuously differentiable. Then*

$$df(t, B_t) = \frac{\partial f}{\partial x}(t, B_t) \, dB_t + \left[\frac{\partial f}{\partial t}(t, B_t) + \frac{1}{2} \frac{\partial^2 f}{\partial x^2}(t, B_t) \right] dt .$$

Proof. (Sort of). We again Taylor expand:

$$\begin{aligned} df &= \frac{\partial f}{\partial x} dB + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} (dB)^2 + \dots \\ &+ \frac{\partial f}{\partial t} dt + \frac{1}{2} \frac{\partial^2 f}{\partial t^2} (dt)^2 + \dots \\ &+ \frac{\partial^2 f}{\partial x \partial t} dB dt + \dots \end{aligned}$$

In the first line, the $(dB)^2$ becomes dt and all higher terms are zero. In the second line, $(dt)^k$ for $k \geq 2$ are zero. And in the last line (mixed partials), all terms have at least one dB and one dt , giving zero. $\square$

Examples.

1. Take $f(x) = x^2$. Then Ito's formula gives

$$B_t^2 - B_0^2 = 2 \int_0^t B_s dB_s + t ,$$

so

$$\int_0^t B_s dB_s = \frac{1}{2} B_t^2 - \frac{t}{2} .$$

In particular, we see that $B_t^2 - t$ is a martingale. Note here that we do not get only $\frac{1}{2} B_t^2$, as we would if we applied the rules of ordinary calculus.

LECTURE 35

We continue with examples of Ito's formula.

1. Take $f(x) = x^k$ for $k \geq 2$. Then

$$B_t^k = k \int_0^t B_s^{k-1} dB_s + \frac{k(k-1)}{2} \int_0^t B_s^{k-2} ds .$$

2. Take $f(x) = e^{\alpha x}$ for $\alpha > 0$. Then

$$e^{\alpha B_t} - 1 = \alpha \int_0^t e^{\alpha B_s} dB_s + \frac{\alpha^2}{2} \int_0^t e^{\alpha B_s} dt .$$

3. (Geometric Brownian motion). Let $X_t = \exp(\sigma B_t + (\mu - \frac{1}{2}\sigma^2)t)$. Then this is $f(t, B_t)$, where

$$f(t, x) = \exp\left(\sigma x + \left(\mu - \frac{1}{2}\sigma^2\right)t\right) .$$

Then

$$\begin{aligned}\frac{\partial f}{\partial x}(t, X_t) &= \sigma X_t \\ \frac{\partial^2 f}{\partial x^2}(t, X_t) &= \sigma^2 X_t \\ \frac{\partial f}{\partial t}(t, X_t) &= (\mu - \frac{1}{2}\sigma^2)X_t .\end{aligned}$$

So by Ito's formula,

$$dX_t = \sigma X_t dB_t + \mu X_t dt .$$

The way to interpret this is that at time t , infinitesimally, X_t changes by a Brownian increment with variance $(\sigma X_t)^2$ and has a drift of μX_t . Since $X_t \geq 0$ for all t , this is a nonnegative drift.

If we take $\mu = 0$ then

$$dX_t = \sigma X_t dB_t .$$

In other words,

$$X_t - 1 = X_t - X_0 = \sigma \int_0^t X_s dB_s ,$$

and we find that X_t is a martingale.

4. Let $X_t = e^{t/2} \cos B_t$. Then $X_t = f(t, B_t)$, where $f(t, x) = e^{t/2} \cos x$. So

$$df = \left[\frac{1}{2}X_t - \frac{1}{2}X_t \right] dt - e^{t/2} \sin B_t dB_t = -e^{t/2} \sin B_t dB_t .$$

So

$$e^{t/2} \cos B_t - 1 = - \int_0^t e^{-s/2} \sin B_s dB_s$$

and X_t is a martingale.

Proof of Ito. Recall the statement: almost surely, the following holds for all $t \in [0, T]$:

$$f(B_t) - f(B_0) = \int_0^t f'(B_s) dB_s + \frac{1}{2} \int_0^t f''(B_s) ds . \quad (33)$$

Let $t_0 \in [0, T]$ be rational and let

$$A_{t_0} = \{(33) \text{ holds for } t = t_0\} .$$

We will show that $\mathbb{P}(A_{t_0}) = 1$. Assuming this, then $A := \cap_{t_0 \in [0, T] \cap \mathbb{Q}} A_{t_0}$ has $\mathbb{P}(A) = 1$ and for $\omega \in A$, (33) holds for all rational $t \in [0, T]$. But both sides of (33) are continuous in t , so on A , it must hold for all real $t \in [0, T]$.

The proof of $\mathbb{P}(A_t) = 1$ is similar to that of the fundamental theorem of calculus, but we use Taylor's theorem instead of the intermediate value theorem, since we must keep track of remainders of order 2. For this reason, define for $\delta > 0$ and $M > 0$

$$\omega(\delta, M) = \sup\{|f''(x) - f''(y)| : x, y \in [-M, M], |x - y| < \delta\}$$

(note that by symmetry the absolute values are not needed above). Let

$$P_n = \{0 = t_0^{(n)} < t_1^{(n)} < \dots < t_{k_n}^{(n)} = t\}, \quad n \geq 1$$

be a deterministic sequence of partitions of $[0, t]$ and, writing $\|P_n\| = \max_{i=1, \dots, k_n} (t_i^{(n)} - t_{i-1}^{(n)})$, assume that $\|P_n\| \rightarrow 0$. On each subinterval $[t_i^{(n)}, t_{i+1}^{(n)}]$ we apply Taylor's formula to the image under B_t : for some c between $x = B_{t_i^{(n)}}$ and $y = B_{t_{i+1}^{(n)}}$,

$$\begin{aligned} f(y) - f(x) - f'(x)(y - x) &= \frac{1}{2}f''(c)(x - y)^2 \\ \left| f(y) - f(x) - f'(x)(y - x) - \frac{1}{2}f''(x)(x - y)^2 \right| &= \frac{1}{2}(x - y)^2 |f''(c) - f''(x)| \\ &\leq \omega(\delta_n, M_t)(x - y)^2, \end{aligned}$$

where $\delta_n = \max_{k=1, \dots, k_n} |B_{t_k^{(n)}} - B_{t_{k-1}^{(n)}}|$ and $M_t = \max_{0 \leq s \leq t} |B_s|$.

Therefore

$$\omega(\delta_n, M_t) \sum_{k=1}^{k_n} (B_{t_k^{(n)}} - B_{t_{k-1}^{(n)}})^2 \geq \left| \sum_{k=1}^{k_n} (f(B_{t_k^{(n)}}) - f(B_{t_{k-1}^{(n)}})) \right. \quad (34)$$

$$\left. - \sum_{k=1}^{k_n} f'(B_{t_{k-1}^{(n)}})(B_{t_k^{(n)}} - B_{t_{k-1}^{(n)}}) \right| \quad (35)$$

$$\left. - \frac{1}{2} \sum_{k=1}^{k_n} f''(B_{t_{k-1}^{(n)}})(B_{t_k^{(n)}} - B_{t_{k-1}^{(n)}})^2 \right|. \quad (36)$$

The term (34) equals $f(B_t) - f(B_0)$. Since f' is continuous, $f'(B_t)$ is a progressively measurable process and by assumption, it is in $\mathcal{M}_t$, so (35) converges to $\int_0^t f'(B_s) dB_s$ in probability. Also f'' is continuous, so a homework problem from this week shows (36) converges in probability to $-\frac{1}{2} \int_0^t f''(B_s) ds$. Choosing a subsequence (n_k) such that all of these convergences occur almost surely, we find

$$\begin{aligned} \left| f(B_t) - f(B_0) - \int_0^t f'(B_s) dB_s - \frac{1}{2} \int_0^t f''(B_s) ds \right| \\ \leq \liminf_k \left[\omega(\delta_{n_k}, M_t) \sum_{i=1}^{n_k-1} (B_{t_{i+1}^{(n_k)}} - B_{t_i^{(n_k)}})^2 \right]. \end{aligned}$$

By the quadratic variation calculation on the homework,

$$\sum_{i=1}^{n_k-1} (B_{t_{i+1}^{(n_k)}} - B_{t_i^{(n_k)}})^2 \rightarrow t \text{ in probability .}$$

So we can find a further subsequence such that this convergence occurs almost surely. Calling this subsequence still (n_k) , the upper bound we get is

$$t \liminf_k \omega(\delta_{n_k}, M_t) .$$

Because f'' is continuous, this is zero almost surely. We conclude that $\mathbb{P}(A_t) = 1$. □

17 Local time

For a simple random walk S_n , we can record the number of times that S_n hits a particular integer x until time N :

$$L_x^N = \sum_{n=1}^N \mathbf{1}_{\{S_n=x\}} .$$

For Brownian motion, the situation is more complicated. Take, for instance, $x = 0$. Then in any interval $[0, T]$, B_t will hit zero infinitely often (actually uncountable many times) and any sum as above will be infinite. So it makes more sense to try to find the amount of time B_t spends at zero. For this purpose, we could consider $\int_0^T \mathbf{1}_{\{B_t=0\}} dt$. Again, this will be zero, since it is just the Lebesgue measure of the zero set. However, we can consider

$$\mu_T(A) := \int_0^T \mathbf{1}_{\{B_t \in A\}} dt \text{ for } A \subset \mathbb{R} \text{ Borel .}$$

Then μ_T is called the *occupation measure* until time T . One way to measure the time at x is to look at the density of the occupation measure at x .

Definition 17.1. *The local time of (B_t) until time T at $x \in \mathbb{R}$ is defined as*

$$L_x^T = \lim_{\epsilon \downarrow 0} \frac{1}{2\epsilon} \int_0^T \mathbf{1}_{\{B_t \in (x-\epsilon, x+\epsilon)\}} dt = \lim_{\epsilon \downarrow 0} \frac{\mu_T(x-\epsilon, x+\epsilon)}{2\epsilon} .$$

We will prove that the limit above exists in probability. Furthermore, using Ito's formula we can find an integral representation, called Tanaka's formula, for it.

17.1 Tanaka's formula

The idea we will pursue is to apply Ito's formula to $|B_t|$. Of course, the function $x \mapsto |x|$ is not twice continuously differentiable. But in the sense of distributional derivatives, we have

$$\frac{d}{dx}|x| = \text{sign}(x) \text{ and } \frac{d^2}{dx^2}|x| = 2\delta_0 .$$

Therefore we might get

$$d|B_t| = \text{sign}(B_t) dB_t + \delta_0(B_t) dt ,$$

or in integral form,

$$\begin{aligned} |B_T| &= \int_0^T \text{sign}(B_t) dB_t + \int_0^T \delta_0(B_t) dt \\ &= \int_0^T \text{sign}(B_t) dB_t + L_0^T , \end{aligned}$$

so that

$$L_0^T = |B_T| - \int_0^T \text{sign}(B_t) dB_t .$$

LECTURE 36

Theorem 17.2 (Tanaka's formula). *For all $T \geq 0$,*

$$L_0^T = \lim_{\epsilon \downarrow 0} \frac{1}{2\epsilon} \mu_T(-\epsilon, \epsilon)$$

exists in probability. Furthermore,

$$L_0^T = |B_T| - \int_0^T \text{sign}(B_t) dB_t \text{ almost surely .}$$

Proof. We cannot apply Ito to the absolute value function but we can after smoothing. For our particular smoothed version, let's take f any smooth function with compact support in $[-2, 2]$, with $0 \leq f \leq 1$ and with $\int f = 1$. Then define the mollified absolute value as

$$f_\epsilon(x) = \epsilon^{-1} \int |x - y| f(y/\epsilon) dy .$$

This is a convolution with a smooth function, so it is smooth. Applying Ito's formula,

$$f_\epsilon(B_T) - f_\epsilon(B_0) = \int_0^T (f_\epsilon)'(B_t) dB_t + \frac{1}{2} \int_0^T (f_\epsilon)''(B_t) dt . \quad (37)$$

To compute the derivatives, do a change of variable $u = x - y$ to rewrite

$$f_\epsilon(x) = \epsilon^{-1} \int |u| f(\epsilon^{-1}(x - u)) du .$$

• So

$$f'_\epsilon(x) = \epsilon^{-1} \lim_{h \rightarrow 0} \int |u| \frac{f(\epsilon^{-1}(x + h - u)) - f(\epsilon^{-1}(x - u))}{h} du .$$

By the mean value theorem, the quotient inside equals

$$\left. \frac{d}{dx} f(\epsilon^{-1}(x - u)) \right|_{x=c} = \epsilon^{-1} f'(\epsilon^{-1}(c - u))$$

for some c between x and $x + h$. Since the derivative of f also has compact support, we can use dominated convergence for

$$f'_\epsilon(x) = \epsilon^{-2} \int |u| f'(\epsilon(x-u)) \, du = \epsilon^{-2} \int_0^\infty u f'(\epsilon^{-1}(x-u)) \, du - \epsilon^{-2} \int_{-\infty}^0 u f'(\epsilon^{-1}(x-u)) \, du .$$

These integrals can be done by parts:

$$\int_0^\infty u f'(\epsilon^{-1}(x-u)) \, du = \epsilon \int_0^\infty f(\epsilon^{-1}(x-u)) \, du$$

and

$$\int_{-\infty}^0 u f'(\epsilon^{-1}(x-u)) \, du = \epsilon \int_{-\infty}^0 f(\epsilon^{-1}(x-u)) \, du ,$$

so

$$f'_\epsilon(x) = \epsilon^{-1} \int \text{sign}(u) f(\epsilon^{-1}(x-u)) \, du , \quad (38)$$

which by a change of variables is

$$f'_\epsilon(x) = \int \text{sign}(x - \epsilon y) f(y) \, dy .$$

- For the second derivative, again we pull the derivative inside (38) for

$$f''_\epsilon(x) = \epsilon^{-2} \int \text{sign}(u) f'(\epsilon^{-1}(x-u)) \, du .$$

This is

$$\epsilon^{-2} \left[\int_0^\infty f'(\epsilon^{-1}(x-u)) \, du - \int_{-\infty}^0 f'(\epsilon^{-1}(x-u)) \, du \right]$$

and we can evaluate using a change of variable for

$$f''_\epsilon(x) = 2\epsilon^{-1} f(\epsilon^{-1}x) .$$

Place these derivatives back into (37) for

$$f_\epsilon(B_T) - f_\epsilon(B_0) = \int_0^T \left[\int \text{sign}(B_t - \epsilon y) f(y) \, dy \right] dB_t + \epsilon^{-1} \int_0^T f(\epsilon^{-1}B_t) \, dt . \quad (39)$$

We will let $\epsilon \downarrow 0$. First for all x ,

$$\begin{aligned} |f_\epsilon(x) - |x|| &= \epsilon^{-1} \left| \int (|x-y| - |x|) f(\epsilon^{-1}y) \, dy \right| \leq \epsilon^{-1} \int ||x-y| - |x|| f(\epsilon^{-1}y) \, dy \\ &= \epsilon^{-1} \int_{-2\epsilon}^{2\epsilon} ||x-y| - |x|| f(\epsilon^{-1}y) \, dy \\ &\leq 2\epsilon . \end{aligned}$$

In other words, $f_\epsilon(x) \rightarrow |x|$ uniformly. Therefore $f_\epsilon(B_T) \rightarrow |B_T|$ and $f_\epsilon(B_0) \rightarrow 0$ almost surely. For the first term on the right of (39),

$$f'_\epsilon(x) = \int \text{sign}(x - \epsilon y) f(y) \, dy \rightarrow \text{sign}(x) \text{ if } x \neq 0$$

and it is bounded by 1. Therefore for each $t > 0$,

$$\mathbb{E}(f'_\epsilon(B_t) - \text{sign}(B_t))^2 \rightarrow 0$$

by the bounded convergence theorem. And again by the bounded convergence theorem and isometry,

$$\mathbb{E} \left[\int_0^T f'_\epsilon(B_t) \, dB_t - \int_0^T \text{sign}(B_t) \, dB_t \right]^2 = \int_0^T \mathbb{E}(f'_\epsilon(B_t) - \text{sign}(B_t))^2 \, dt \rightarrow 0 .$$

This means $\int_0^T f'_\epsilon(B_t) \, dB_t \rightarrow \int_0^T \text{sign}(B_t) \, dB_t$ in L^2 and therefore in probability.

Coming back to (39), we have shown that

$$\lim_{\epsilon \downarrow 0} \epsilon^{-1} \int_0^T f(\epsilon^{-1} B_t) \, dt = |B_T| - \int_0^T \text{sign}(B_t) \, dB_t \text{ in probability .} \quad (40)$$

Now we want to choose f to be close to $(1/2)\mathbf{1}_{(-1,1)}$. Given $\delta > 0$ pick f to be smooth such that f has support in $[-2, 2]$, $f \geq (1/2)\mathbf{1}_{(-1,1)}$ and $\int f = 1 + \delta$. Then $f/(1 + \delta)$ satisfies the conditions of the above derivation and (40) holds for $f/(1 + \delta)$. This means

$$\begin{aligned} \frac{1}{2\epsilon} \int_0^T \mathbf{1}_{(-\epsilon, \epsilon)}(B_t) \, dt &\leq \epsilon^{-1} \int_0^T f(\epsilon^{-1} B_t) \, dt \\ &\rightarrow (1 + \delta) \left[|B_T| - \int_0^T \text{sign}(B_t) \, dB_t \right] \text{ in probability .} \end{aligned}$$

We will finish up next time. □

LECTURE 37

We were almost done proving:

Theorem 17.3 (Tanaka's formula). *For all $T \geq 0$,*

$$L_0^T = \lim_{\epsilon \downarrow 0} \frac{1}{2\epsilon} \mu_T(-\epsilon, \epsilon)$$

exists in probability. Furthermore,

$$L_0^T = |B_T| - \int_0^T \text{sign}(B_t) \, dB_t \text{ almost surely .}$$

Proof. We have shown that for any $\delta > 0$,

$$\begin{aligned} \frac{1}{2\epsilon} \int_0^T \mathbf{1}_{(-\epsilon, \epsilon)}(B_t) dt &\leq \epsilon^{-1} \int_0^T f(\epsilon^{-1} B_t) dt \\ &\rightarrow (1 + \delta) \left[|B_T| - \int_0^T \text{sign}(B_t) dB_t \right] \text{ in probability ,} \end{aligned}$$

where f is a function that dominates $\frac{1}{2}\mathbf{1}_{(-\epsilon, \epsilon)}$. So

$$\lim_{\epsilon \downarrow 0} \mathbb{P} \left(\frac{1}{2\epsilon} \int_0^T \mathbf{1}_{(-\epsilon, \epsilon)}(B_t) dt \geq (1 + \delta) \left[|B_T| - \int_0^T \text{sign}(B_t) dB_t \right] + \delta \right) = 0 .$$

This is the upper bound (in probability) in Tanaka's formula. For the lower bound, take f with support in $[-2, 2]$, $0 \leq f \leq (1/2)\mathbf{1}_{(-1, 1)}$ and $\int f = 1 - \delta$ and repeat the above bound with $\geq$ in place of $\leq$. $\square$

Some remarks:

- A similar argument gives existence of local time at other x -values:

$$L_x^T = |B_T - x| - |x| - \int_0^T \text{sign}(B_s - x) dB_s \text{ almost surely .}$$

- We can easily get a different representation:

$$\frac{1}{2} L_x^T = (B_T - x)_+ - x_- - \int_0^T \mathbf{1}_{\{B_s > x\}} dB_s$$

or

$$\frac{1}{2} L_x^T = (B_T - x)_- - x_+ + \int_0^T \mathbf{1}_{\{B_s \leq x\}} dB_s .$$

Proof. The sum of the two right sides is

$$|B_T - x| - |x| - \int_0^T \text{sign}(B_s - x) dB_s ,$$

which is L_x^T by Tanaka. On the other hand, the difference is

$$(B_T - x)_+ - (B_T - x)_- - x_- + x_+ - \int_0^T 1 dB_s ,$$

which is

$$B_T - x + x - B_T = 0 .$$

This means they are both equal to $(1/2)L_x^T$. $\square$

- For each x , the function $T \mapsto L_x^T$ has a continuous version which is non-decreasing and is constant on each component of the complement of the set $\{t : B_t = x\}$.

Proof. Continuity follows from the integral representation. Non-decreasing is a homework problem.

For the final statement, let $r < s$ be rationals and use the last representation for

$$L_x^s - L_x^r = 2 \left[(B_s - x)_+ - (B_r - x)_+ - \int_r^s \mathbf{1}_{\{B_u > x\}} dB_u \right] .$$

On the event that the Brownian motion stays to the right of x for the time interval $[r, s]$,

$$(B_s - x)_+ = B_s - x, \quad (B_r - x)_+ = B_r - x \text{ and } \int_r^s \mathbf{1}_{\{B_u > x\}} dB_u = B_r - B_s ,$$

so $L_x^s = L_x^r$ on this event.

Now if $L_x^{t_1} \neq L_x^{t_2}$ for two times $t_1 < t_2$, both in the same component of the complement of $\{t : B_t = x\}$ (note this is a closed set, so its complement is a union of open intervals) then by continuity of local time, we can find two rationals $r < s$ in this same component such that $L_x^r \neq L_x^s$. Without loss in generality, B_t stays to the right of x on this interval and as we have seen above, this has probability zero. $\square$

17.2 Distribution of local time

We will now analyze the formula $L_0^T = |B_T| - \int_0^T \text{sign}(B_s) dB_s$ to find the distribution of local time.

Proposition 17.4. *The process $X_t = \int_0^t \text{sign}(B_s) dB_s$ is a Brownian motion.*

Proof. For this result we need a lemma that we will prove next week.

Lemma 17.5 (Levy's characterization of Brownian motion). *Let (X_t) be a martingale with continuous paths and $X_0 = 0$ such that $X_t^2 - t$ is a martingale. Then (X_t) is a Brownian motion.*

Our X_t is a continuous martingale, so we verify the second condition. For $0 \leq s \leq t$,

$$\mathbb{E}[X_t^2 - X_s^2 \mid \mathcal{F}_s] = \mathbb{E}[(X_t - X_s)^2 \mid \mathcal{F}_s] + 2\mathbb{E}[X_s(X_t - X_s) \mid \mathcal{F}_s] .$$

The second term is

$$2X_s\mathbb{E}[X_t - X_s \mid \mathcal{F}_s] = 0 ,$$

since (X_t) is a martingale. The first term is

$$\mathbb{E} \left[\left(\int_s^t \text{sign}(B_u) dB_u \right)^2 \mid \mathcal{F}_s \right]$$

and can be evaluated by approximating with simple processes. To this end, for each n , let P_n be a partition of $[s, t]$:

$$P_n = \{s = t_0^{(n)} < \dots < t_{k_n}^{(n)} = t\}$$

and define the simple process $S_u^{(n)}$ by

$$S^{(n)} = \sum_{k=1}^{k_n} \text{sign}(B_{t_{k-1}^{(n)}}) \mathbf{1}_{(t_{k-1}^{(n)}, t_k^{(n)}]} .$$

Now

$$\int_s^t S_u^{(n)} dB_u = \sum_{k=1}^{k_n} y_k ,$$

where $y_k = \text{sign}(B_{t_{k-1}^{(n)}})(B_{t_k^{(n)}} - B_{t_{k-1}^{(n)}})$. As we have seen many times, for $j < k$,

$$\mathbb{E}[y_j y_k \mid \mathcal{F}_s] = \mathbb{E} \left[y_k \mathbb{E}[y_k \mid \mathcal{F}_{t_{k-1}^{(n)}}] \mid \mathcal{F}_s \right] = 0 ,$$

so

$$\mathbb{E} \left[\left(\int_s^t S_u^{(n)} dB_u \right)^2 \mid \mathcal{F}_s \right] = \sum_{k=1}^{k_n} \mathbb{E}[y_k^2 \mid \mathcal{F}_s] = \sum_{k=1}^{k_n} \mathbb{E} \left[(B_{t_k^{(n)}} - B_{t_{k-1}^{(n)}})^2 \mid \mathcal{F}_s \right] = t - s .$$

For notational convenience, set $A = \int_s^t \text{sign}(B_u) dB_u$ and $A_n = \int_s^t S_u^{(n)} dB_u$, so that $A_n \rightarrow A$ in L^2 . Therefore by Jensen and Cauchy-Schwarz,

$$\mathbb{E} \left| \mathbb{E}[A_n^2 \mid \mathcal{F}_s] - \mathbb{E}[A^2 \mid \mathcal{F}_s] \right| \leq \mathbb{E} |A_n^2 - A^2| \leq \|A_n - A\|_2 \|A_n + A\|_2 \rightarrow 0 .$$

The left side is $\mathbb{E} |\mathbb{E}[A^2 \mid \mathcal{F}_s] - (t - s)|$, so $\mathbb{E}[A^2 \mid \mathcal{F}_s] = t - s$ almost surely. Returning to the beginning,

$$\mathbb{E}[X_t^2 - X_s^2 \mid \mathcal{F}_s] = t - s ,$$

or

$$\mathbb{E}[(X_t^2 - t) - (X_s^2 - s) \mid \mathcal{F}_s] = 0 .$$

□

The above proof can be used to show more: if (X_t) is in $\mathcal{M}_T$ then the process

$$\left(\int_0^t X_s dB_s \right)^2 - \int_0^t X_s^2 ds$$

is a martingale. In particular, by Levy's theorem, if (X_t) is ± 1 -valued and progressively measurable, then $\int_0^t X_s dB_s$ is a Brownian motion.

From this proposition we can give a different characterization of local time.

Theorem 17.6 (Levy's identity for local time). *Let $M_t = \max_{0 \leq s \leq t} B_s$. Then the pairs of processes*

$$(|B_t|, L_0^t) \text{ and } (M_t - B_t, M_t)$$

have the same distribution.

Before we prove this theorem let us think about what it says about the structure of local time.

- First, for each fixed t , L_0^t has the same distribution as M_t . Therefore for each $a \geq 0$,

$$\mathbb{P}(L_0^t \geq a) = 2\mathbb{P}(B_t \geq a) .$$

In particular, for each $t \geq 0$, L_0^t has zero probability to be 0. That is, in every nonzero interval $[0, t]$, Brownian motion accumulates non-trivial local time.

- The entire process L_0^t has the same distribution as the process M_t . The maximum process M_t is clearly continuous, non-decreasing and only increases when Brownian motion achieves a new maximum (think about the structure of the set of times on which B_t achieves a new max). We can imagine drawing M_t as a function of time pretty easily, so this gives us a way to visualize the process L_0^t .
- The process M_t increases when $M_t - B_t$ comes to 0. Since this pair has the same distribution as $(|B_t|, L_0^t)$, this means that L_0^t increases when $|B_t|$ comes to 0.
- Last, $|B_t|$ has the same distribution as $M_t - B_t$. The first process is sometimes called reflected Brownian motion.

LECTURE 38

The first goal today is to prove:

Theorem 17.7 (Levy's identity for local time). *Let $M_t = \max_{0 \leq s \leq t} B_s$. Then the pairs of processes*

$$(|B_t|, L_0^t) \text{ and } (M_t - B_t, M_t)$$

have the same distribution.

Proof. The previous proposition, along with Tanaka's formula, gives $L_0^t + W_t = |B_t|$, where W_t is a Brownian motion:

$$W_t = \int_0^t \text{sign}(B_s) dB_s .$$

Define the Brownian motion $Y_t = -W_t$ and its maximum process $N_t = \max_{0 \leq s \leq t} Y_s$. We claim that

$$N_t = L_0^t \text{ for all } t \geq 0 .$$

To show this, we prove both inequalities. First, for all s ,

$$Y_s = -W_s = L_0^s - |B_s| \leq L_0^s$$

so $N_t = \max_{0 \leq s \leq t} Y_s \leq \max_{0 \leq s \leq t} L_0^s = L_0^t$ (since L_0^t is nondecreasing). For the other inequality, $L_0^t \leq N_t$, consider two cases. If $B_t = 0$ then

$$N_t \geq Y_t = L_0^t - |B_t| = L_0^t .$$

Otherwise, t is in some component of the complement of $\{s : B_s = 0\}$ and L_0^s is constant on this interval. Therefore setting t_0 to be the left endpoint of this interval, $B_{t_0} = 0$ and so

$$L_0^t = L_0^{t_0} = Y_{t_0} \leq N_{t_0} \leq N_t .$$

This establishes the claim.

Since $N_t = L_0^t$ for all $t \geq 0$ and $Y_t = L_0^t - |B_t|$, we must have

$$|B_t| = L_0^t - Y_t = N_t - Y_t .$$

So

$$(|B_t|, L_t^0) = (N_t - Y_t, N_t) .$$

The last process has the same distribution as $(M_t - B_t, M_t)$.

□

18 Levy's characterization of Brownian motion

To prove the last result we used:

Theorem 18.1 (Levy's characterization of Brownian motion). *Let (X_t) be a martingale with continuous paths and $X_0 = 0$ such that $X_t^2 - t$ is a martingale. Then (X_t) is a Brownian motion.*

Proof. The proof is taken from Varadhan's stochastic processes book and will be in two steps. The bulk of the proof is to show that for each $\lambda \in \mathbb{R}$,

$$Y_t = Y_t^{(\lambda)} := \exp \left(\lambda X_t - \frac{\lambda^2 t}{2} \right) \tag{41}$$

is a martingale. Let us assume this for the moment and show that it suffices to prove the theorem.

Step 1. If (43) holds, then let $0 \leq s \leq t$. We obtain

$$\begin{aligned} \mathbb{E} [\exp (\lambda(X_t - X_s)) \mid \mathcal{F}_s] &= e^{-\lambda X_s} e^{\frac{\lambda^2 t}{2}} \mathbb{E} \left[\exp \left(\lambda X_t - \frac{\lambda^2 t}{2} \right) \mid \mathcal{F}_s \right] \\ &= e^{-\lambda X_s} e^{\frac{\lambda^2 t}{2}} e^{\lambda X_s - \frac{\lambda^2 s}{2}} \\ &= \exp \left(\frac{\lambda^2 (t - s)}{2} \right) . \end{aligned}$$

If $0 \leq t_0 \leq \dots \leq t_k$, we want to use the above property to show that $(Y_1, \dots, Y_k) := (X_{t_1} - X_{t_0}, \dots, X_{t_k} - X_{t_{k-1}})$ is Gaussian with independent entries of variance $t_i - t_{i-1}$. To do this, let's use the Cramer-Wold device. Letting $(G_1, \dots, G_k)$ be a Gaussian vector with independent entries and variances $t_1 - t_0, \dots, t_k - t_{k-1}$, we must show that for each vector of numbers $(\lambda_1, \dots, \lambda_k)$, the variables

$$\lambda_1 Y_1 + \dots + \lambda_k Y_k \text{ and } \lambda_1 G_1 + \dots + \lambda_k G_k$$

have the same distribution. The variable on the right is Gaussian with mean zero and variance $\sum_{i=1}^k \lambda_i(t_i - t_{i-1})$, so its characteristic function is

$$\exp\left(-\frac{t^2 \sum_{i=1}^k \lambda_i(t_i - t_{i-1})}{2}\right). \quad (42)$$

To show that the left variable has this characteristic function, we find its moment generating function first.

By applying the Tower property k times, along with the previous calculation (conditioning on $\mathcal{F}_{t_{k-1}}, \dots, \mathcal{F}_{t_0}$ successively), for $t \in \mathbb{R}$,

$$\begin{aligned} \mathbb{E} e^{t(\lambda_1 Y_1 + \dots + \lambda_k Y_k)} &= \mathbb{E} \left[e^{t\lambda_k(X_{t_k} - X_{t_{k-1}})} \dots e^{t\lambda_1(X_{t_1} - X_{t_0})} \right] = \prod_{i=1}^k \exp\left(\frac{t^2 \lambda_i^2(t_i - t_{i-1})}{2}\right) \\ &= \exp\left(\frac{t^2 \sum_{i=1}^k \lambda_i^2(t_i - t_{i-1})}{2}\right). \end{aligned}$$

This is the moment generating function of $\lambda_1 G_1 + \dots + \lambda_k G_k$, and we can take its derivatives at $t = 0$ to find that all moments agree as well. Last, we can plug these moments into the power series expansion for the characteristic function (since they do not grow too quickly) to find that $\lambda_1 Y_1 + \dots + \lambda_k Y_k$ also has characteristic function (42).

Next time we will prove (43). □

LECTURE 39

We continue with the proof of:

Theorem 18.2 (Levy's characterization of Brownian motion). *Let (X_t) be a martingale with continuous paths and $X_0 = 0$ such that $X_t^2 - t$ is a martingale. Then (X_t) is a Brownian motion.*

Proof. We now prove that for each $\lambda \in \mathbb{R}$,

$$Y_t = Y_t^{(\lambda)} := \exp\left(\lambda X_t - \frac{\lambda^2 t}{2}\right) \quad (43)$$

is a martingale.

The proof is actually like the proof of a central limit theorem. In other words, we will write X_t as a sum of many small increments to prove that it is Gaussian distributed. So our first goal is to show

$$\mathbb{E} \exp \left(\lambda X_t - \frac{\lambda^2 t}{2} \right) = 1 \text{ for } \lambda \in \mathbb{R}, t \geq 0. \quad (44)$$

In fact, this will actually suffice for the proof. The reason is that the martingale equality we want to show for $0 \leq s \leq t$ is

$$\mathbb{E} \left[\exp \left(\lambda X_t - \frac{\lambda^2 t}{2} \right) \mid \mathcal{F}_s \right] = \exp \left(\lambda X_s - \frac{\lambda^2 s}{2} \right),$$

or

$$\mathbb{E} \left[\exp \left(\lambda (X_t - X_s) - \frac{\lambda^2}{2} (t - s) \right) \mid \mathcal{F}_s \right] = 1 \text{ almost surely.}$$

The process $(X_{s+t} - X_s)_{t \geq 0}$ is also a continuous martingale starting from 0 such that $(X_{s+t} - X_s)^2 - t$ is a martingale. Therefore it satisfies the assumptions of this theorem. You can check that the proof of (44) below also works for this process when we replace $\mathbb{E}[\cdot]$ with $\mathbb{E}[\cdot \mid \mathcal{F}_s]$. (We will remark this later again when you can check easily.)

Step 2. (Upper bound) To this end, we define a sequence of stopping times for each $\epsilon > 0$:

$$\tau_{0,\epsilon} = 0 \text{ and } \tau_{k+1,\epsilon} = \min \left\{ t, \tau_{k,\epsilon} + \epsilon, \inf \{ s : s \geq \tau_{k,\epsilon}, |X_s - X_{\tau_{k,\epsilon}}| \geq \epsilon \} \right\}.$$

We will use these times to estimate X_t , so we want them to be no bigger than t ; this is the reason for the first term. We also want many of them; that is, we want them to be not very far apart. This is the reason for the second term. And the third term is there to make sure that the increments of X are not large.

Note that each $\tau_{k,\epsilon}$ is a stopping time and $\tau_{k,\epsilon} = t$ for all large k . Therefore we can write

$$X_t = \sum_{k=0}^{\infty} (X_{\tau_{k+1,\epsilon}} - X_{\tau_{k,\epsilon}}) \text{ and } t = \sum_{k=0}^{\infty} (\tau_{k+1,\epsilon} - \tau_{k,\epsilon}).$$

To analyze the k -th increment, set $z = X_{\tau_{k+1,\epsilon}} - X_{\tau_{k,\epsilon}}$ and $\tau = \tau_{k+1,\epsilon} - \tau_{k,\epsilon}$. Then, setting $\mathcal{F}_k = \mathcal{F}_{\tau_{k,\epsilon}}$,

$$\mathbb{E}[z \mid \mathcal{F}_k] = \mathbb{E}[X_{\tau_{k+1,\epsilon}} \mid \mathcal{F}_k] - X_{\tau_{k,\epsilon}} = 0$$

and

$$\begin{aligned} \mathbb{E}[z^2 \mid \mathcal{F}_k] &= \mathbb{E}[X_{\tau_{k+1,\epsilon}}^2 \mid \mathcal{F}_k] - 2X_{\tau_{k,\epsilon}} \mathbb{E}[X_{\tau_{k+1,\epsilon}} - X_{\tau_{k,\epsilon}} \mid \mathcal{F}_k] - X_{\tau_{k,\epsilon}}^2 \\ &= \mathbb{E}[X_{\tau_{k+1,\epsilon}}^2 - \tau_{k+1,\epsilon} \mid \mathcal{F}_k] + \mathbb{E}[\tau_{k+1,\epsilon} \mid \mathcal{F}_k] - X_{\tau_{k,\epsilon}}^2 \\ &= X_{\tau_{k,\epsilon}}^2 - \tau_{k,\epsilon} + \mathbb{E}[\tau_{k+1,\epsilon} \mid \mathcal{F}_k] - X_{\tau_{k,\epsilon}}^2 \\ &= \mathbb{E}[\tau_{k+1,\epsilon} - \tau_{k,\epsilon} \mid \mathcal{F}_k] \\ &= \mathbb{E}[\tau \mid \mathcal{F}_k]. \end{aligned}$$

This gives

$$\mathbb{E}[z^2 - \tau \mid \mathcal{F}_k] = 0 .$$

Therefore, using $|z|, \tau \leq \epsilon$, if $\delta \in [0, 1]$,

$$\mathbb{E} \left[\exp \left(\lambda z - \left(\frac{\lambda^2}{2} + \delta \right) \tau \right) \mid \mathcal{F}_k \right]$$

is bounded by

$$\mathbb{E} \left[1 + \left(\lambda z - \left(\frac{\lambda^2}{2} + \delta \right) \tau \right) + \frac{1}{2} \left(\lambda z + \left(\frac{\lambda^2}{2} + \delta \right) \tau \right)^2 + C_\lambda(|z|^3 + \tau^3) \mid \mathcal{F}_k \right]$$

This is evaluated as

$$1 + \mathbb{E} \left[-\delta\tau + \lambda z \left(\frac{\lambda^2}{2} + \delta \right) \tau + \frac{1}{2} \left(\frac{\lambda^2}{2} + \delta \right)^2 \tau^2 + C_\lambda(|z|^3 + \tau^3) \mid \mathcal{F}_k \right] ,$$

which is bounded by

$$\begin{aligned} 1 + \mathbb{E} \left[-\delta\tau + |\lambda|\epsilon \left(\frac{\lambda^2}{2} + \delta \right) \tau + \frac{\epsilon}{2} \left(\frac{\lambda^2}{2} + \delta \right)^2 \tau + C_\lambda(\epsilon z^2 + \epsilon^2 \tau) \mid \mathcal{F}_k \right] \\ \leq 1 + \mathbb{E}[(-\delta + C'_\lambda \epsilon)\tau \mid \mathcal{F}_k] \end{aligned}$$

for some other constant C'_λ . If $\delta = C'_\lambda \epsilon$, we obtain

$$\mathbb{E} \left[\exp \left(\lambda z - \left(\frac{\lambda^2}{2} + \delta \right) \tau \right) \mid \mathcal{F}_k \right] \leq 1 \text{ almost surely .}$$

We will continue next time. □

LECTURE 40

Proof. We finished last time by showing

$$\mathbb{E} \left[\exp \left(\lambda z - \left(\frac{\lambda^2}{2} + \delta \right) \tau \right) \mid \mathcal{F}_k \right] \leq 1 \text{ almost surely .}$$

For this same choice of δ , we can successively condition on $\mathcal{F}_n, \mathcal{F}_{n-1}, \dots$ to get

$$\mathbb{E} \exp \left(\lambda \sum_{k=1}^n (X_{\tau_{k+1,\epsilon}} - X_{\tau_{k,\epsilon}}) - \left(\frac{\lambda^2}{2} + \delta \right) \sum_{k=1}^n (\tau_{k+1,\epsilon} - \tau_{k,\epsilon}) \right) \leq 1 .$$

By Fatou,

$$\mathbb{E} \exp \left(\lambda X_t - \left(\frac{\lambda^2}{2} + \delta \right) t \right) \leq 1 .$$

Taking $\epsilon \downarrow 0$, also $\delta \downarrow 0$, so

$$\mathbb{E} \exp \left(\lambda X_t - \frac{\lambda^2 t}{2} \right) \leq 1 .$$

In the derivation above, we can replace $\mathbb{E}[\cdot]$ with $\mathbb{E}[\cdot \mid \mathcal{F}_s]$ (and consider the process $X_{t+s} - X_s$), as described in last lecture. By then repeating exactly the same argument, we obtain

$$\mathbb{E} \left[\exp \left(\lambda(X_t - X_s) - \frac{\lambda^2}{2}(t - s) \right) \mid \mathcal{F}_s \right] \leq 1 .$$

This statement has two consequences: $\exp \left(\lambda X_t - \frac{\lambda^2 t}{2} \right)$ is integrable, and it is a supermartingale.

Step 3. (Lower bound) For the other inequality, essentially the same calculation as above shows

$$\mathbb{E} \left[\exp \left(\lambda z - \left(\frac{\lambda^2}{2} - \delta \right) \tau \right) \mid \mathcal{F}_k \right] \geq 1 + \mathbb{E}[(\delta - C''_{\lambda} \epsilon) \tau \mid \mathcal{F}_k] ,$$

and so by successive conditioning,

$$\mathbb{E} \exp \left(\lambda \sum_{k=1}^n (X_{\tau_{k+1,\epsilon}} - X_{\tau_{k,\epsilon}}) - \left(\frac{\lambda^2}{2} - \delta \right) \sum_{k=1}^n (\tau_{k+1,\epsilon} - \tau_{k,\epsilon}) \right) \geq 1 .$$

To take the limit in n , we will need to use dominated convergence. For this, we need a lemma.

Lemma 18.3. *Let (Z_t) be a nonnegative supermartingale with continuous paths. Then for each $a > 0$,*

$$\mathbb{P} \left(\max_{0 \leq s \leq t} Z_t \geq a \right) \leq \frac{\mathbb{E} Z_0}{a} .$$

Proof. Just as in the martingale inequality, we first use a discrete version. Suppose that $Z_1, Z_2, \dots$ is a nonnegative supermartingale and define the stopping time

$$\tau = \min\{n \geq 1 : Z_n \geq a\} .$$

Then by the optional stopping theorem,

$$\mathbb{E} Z_1 \geq \mathbb{E} Z_{\min\{\tau, n\}} \geq \mathbb{E} Z_{\tau} \mathbf{1}_{\{\tau \leq n\}} \geq a \mathbb{P}(\tau \leq n) .$$

But $\{\tau \leq n\} = \{\max_{1 \leq j \leq n} Z_j \geq a\}$, so

$$\mathbb{P} \left(\max_{1 \leq j \leq n} Z_j \geq a \right) \leq \frac{\mathbb{E} Z_1}{a} .$$

Now apply this result to the supermartingale $Z_0, Z_{t2^{-n}}, Z_{2t2^{-n}}, \dots, Z_t$ to obtain

$$\mathbb{P} \left(\max_{0 \leq j \leq 2^n} Z_{jt2^{-n}} \geq a \right) \leq \frac{\mathbb{E} Z_0}{a} .$$

By continuity, as $n \rightarrow \infty$, the probability on the left converges to $\mathbb{P}(\max_{0 \leq s \leq t} Z_t \geq a)$, completing the proof. $\square$

Apply the lemma to our supermartingale for

$$\mathbb{P} \left(\max_{0 \leq s \leq t} \exp \left(\lambda X_s - \frac{\lambda^2 s}{2} \right) \geq a \right) \leq \frac{1}{a} .$$

This is true for all λ , so we can use it to get a tail estimate for the maximum of X_s . For $\lambda > 0$,

$$\begin{aligned} \mathbb{P} \left(\max_{0 \leq s \leq t} X_s \geq a \right) &\leq \mathbb{P} \left(\max_{0 \leq s \leq t} \exp \left(\lambda X_s - \frac{\lambda^2 s}{2} \right) \geq \exp \left(\lambda a - \frac{\lambda^2 t}{2} \right) \right) \\ &\leq \exp \left(-\lambda a + \frac{\lambda^2 t}{2} \right) . \end{aligned}$$

Choose $\lambda = a/t$ for the bound

$$\mathbb{P} \left(\max_{0 \leq s \leq t} X_s \geq a \right) \leq \exp \left(-\frac{a^2}{2t} \right) .$$

Therefore for $\lambda \geq 0$, if we set $M_t = \max_{0 \leq s \leq t} X_s$,

$$\mathbb{E} \max_{0 \leq s \leq t} e^{\lambda X_s} = \mathbb{E} e^{\lambda M_t} \leq \sum_{n=0}^{\infty} e^{\lambda n} \mathbb{P}(M_t \geq n) \leq \sum_{n=0}^{\infty} e^{\lambda n} e^{-\frac{n^2}{2t}} < \infty .$$

This means that

$$\max_{0 \leq s \leq t} \exp(\lambda X_s + \delta t)$$

is integrable, and since it is an upper bound for

$$\exp \exp \left(\lambda \sum_{k=1}^n (X_{\tau_{k+1,\epsilon}} - X_{\tau_{k,\epsilon}}) - \left(\frac{\lambda^2}{2} - \delta \right) \sum_{k=1}^n (\tau_{k+1,\epsilon} - \tau_{k,\epsilon}) \right) \leq e^{\lambda X_{\tau_{n+1,\epsilon}}} ,$$

we can apply dominated convergence for

$$\mathbb{E} \exp \left(\lambda X_t - \left(\frac{\lambda^2}{2} - \delta \right) t \right) \geq 1 .$$

Again we take $\delta \rightarrow 0$ with ϵ to obtain

$$\mathbb{E} \exp \left(\lambda X_t - \frac{\lambda^2 t}{2} \right) \geq 1 .$$

Because we can replace X_t with $X_t - X_s$ and $\mathbb{E}[\cdot]$ with $\mathbb{E}[\cdot \mid \mathcal{F}_s]$, we have shown the other half of the martingale condition:

$$\mathbb{E} \left[\exp \left(\lambda X_t - \frac{\lambda^2 t}{2} \right) \mid \mathcal{F}_s \right] \geq 1$$

and we are done. □

19 Stationary processes

In this final part of the course, we will describe stationary sequences of random variables. We introduced martingales as a generalization of i.i.d. sequences. Here we will generalize i.i.d. sequences in a different way.

Definition 19.1. *A sequence (X_n) is stationary if for each k and choices $1 \leq i_1 < \dots < i_k$, the vectors $(X_{i_1}, \dots, X_{i_k})$ and $(X_{i_1+1}, \dots, X_{i_k+1})$ have the same distribution.*

Here are some remarks about the definition.

- Of course $(X_{i_1}, \dots, X_{i_k})$ and $(X_{i_1+m}, \dots, X_{i_k+m})$ have the same distribution for $m \geq 1$ if the sequence is stationary.
- Stationarity implies that the X_n 's are identically distributed (take $k = 1$).

LECTURE 41

More remarks about stationary sequences.

- A simple example comes from measure preserving transformations. Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space and $T : \Omega \rightarrow \Omega$ measurable satisfying $\mathbb{P} \circ T^{-1} = \mathbb{P}$. Recall this means

$$\mathbb{P}(A) = \mathbb{P}(T^{-1}A) \text{ for all } A \in \Sigma .$$

Such a T we call *measure preserving*. Let $X : \Omega \rightarrow \mathbb{R}$ be a random variable and define a sequence $X_1, X_2, \dots$ by

$$X_1(\omega) = X(\omega) \text{ and } X_{k+1}(\omega) = X_k(T(\omega)) \text{ for } k \geq 1 .$$

We claim that (X_n) is stationary. Indeed, if $k \geq 1$ and $1 \leq i_1 < \dots < i_k$ then for any Borel $A \subset \mathbb{R}^k$,

$$T^{-1}\{\omega : (X_{i_1}(\omega), \dots, X_{i_k}(\omega)) \in A\} = \{\omega : (X_{i_1+1}(\omega), \dots, X_{i_k+1}(\omega)) \in A\} .$$

Since T is measure preserving,

$$\mathbb{P}((X_{i_1}, \dots, X_{i_k}) \in A) = \mathbb{P}((X_{i_1+1}, \dots, X_{i_k+1}) \in A) ,$$

so these vectors have the same distribution.

- Conversely, if (X_n) is stationary on some probability space $(\Omega, \Sigma, \mathbb{P})$, then we can exhibit the sequence as the last example. In other words, we can find a space $\hat{\Omega}$, a probability measure $\hat{\mathbb{P}}$, a $\hat{\mathbb{P}}$ -preserving measurable map $T : \hat{\Omega} \rightarrow \hat{\Omega}$ and a variable $X : \hat{\Omega} \rightarrow \mathbb{R}$ such that $(X, X(T(\omega)), X(T^2(\omega)), \dots)$ has the same distribution as (X_n) . Let $\hat{\Omega}$ be the space $\mathbb{R}^{\mathbb{N}}$ with the product sigma-algebra. Then define $\Phi : \Omega \rightarrow \hat{\Omega}$ by

$$\Phi(\omega) = (X_1(\omega), X_2(\omega), \dots)$$

and let $\hat{\mathbb{P}}$ be the push forward $\mathbb{P} \circ \Phi^{-1}$. Last let $T : \hat{\Omega} \rightarrow \hat{\Omega}$ be the shift map

$$T(\omega_1, \omega_2, \dots) = (\omega_2, \omega_3, \dots)$$

and define the random variable $X : \hat{\Omega} \rightarrow \mathbb{R}$ by

$$X(\omega_1, \omega_2, \dots) = \omega_1 .$$

Because Φ is measurable (you showed this on homework), $\hat{\mathbb{P}}$ is T -invariant and the sequence

$$(X, X \circ T, X \circ T^2, \dots) = (\omega_1, \omega_2, \dots)$$

has the same distribution as $(X_1, X_2, \dots)$.

So we may view our stationary sequence as an element of the sequence space $\mathbb{R}^{\mathbb{N}}$ along with the measure preserving shift map T . So we will from now adopt this point of view, common in dynamical systems, and consider measure preserving systems.

Given $(\Omega, \Sigma, \mathbb{P})$ along with a measure-preserving (measurable) map $T : \Omega \rightarrow \Omega$, we may begin by asking if we can decompose our system into smaller parts. One example comes from invariant events. Suppose $A \in \Sigma$ satisfies $T^{-1}A = A$. Then we can restrict our system to A .

Proposition 19.2. *Let $\mathcal{I}$ be the collection of all T -invariant events:*

$$\mathcal{I} = \{A \in \Sigma : T^{-1}A = A\} .$$

Then $\mathcal{I}$ is a sigma-algebra, the invariant sigma-algebra. If $A \in \mathcal{I}$ and $\mathbb{P}(A) > 0$ then T is invariant for the conditioned measure $\mathbb{P}(\cdot | A)$.

Proof. The proof is pretty easy. The events Ω is in $\mathcal{I}$ and if $A \in \mathcal{I}$ then

$$T^{-1}A^c = (T^{-1}A)^c = A^c ,$$

so $A^c \in \mathcal{I}$. Last if $A_1, A_2, \dots \in \mathcal{I}$ then

$$T^{-1}(\cup_n A_n) = \cup_n T^{-1}A_n = \cup_n A_n$$

and this shows $\mathcal{I}$ is a sigma-algebra.

If $A \in \mathcal{I}$ has $\mathbb{P}(A) > 0$ then for $B \in \Sigma$,

$$\begin{aligned} \mathbb{P}(T^{-1}B | A) &= \frac{\mathbb{P}((T^{-1}B) \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}((T^{-1}B) \cap (T^{-1}A))}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(T^{-1}(B \cap A))}{\mathbb{P}(A)} = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \mathbb{P}(B | A) . \end{aligned}$$

□

In view of this proposition, the simplest systems will have $\mathcal{I}$ trivial. In a sense they are irreducible.

Definition 19.3. A measurable $T : \Omega \rightarrow \Omega$ is called ergodic for $(\Omega, \Sigma, \mathbb{P})$ if

$$A \in \mathcal{I} \Rightarrow \mathbb{P}(A) = 0 \text{ or } 1 .$$

The notation is a bit unclear in the literature. Sometimes the system $(\Omega, \Sigma, \mathbb{P}, T)$ is called ergodic. Sometimes T is ergodic for $\mathbb{P}$. Sometimes $\mathbb{P}$ is ergodic for T . We will use these interchangeably.

There are various alternative characterizations of ergodicity. Here is one:

Proposition 19.4. T is ergodic for $\mathbb{P}$ if and only if every T -invariant measurable function $f : \Omega \rightarrow \mathbb{R}$ is constant almost surely.

Proof. If every T -invariant function is constant, take $A \in \mathcal{I}$ and consider $f = \mathbf{1}_A$. This is T -invariant:

$$f(T(\omega)) = \mathbf{1}_A(T(\omega)) = \begin{cases} 1 & \text{if } T(\omega) \in A \\ 0 & \text{if } T(\omega) \notin A \end{cases} = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} = f(\omega) .$$

So $\mathbf{1}_A$ is constant almost surely and must then be 0 or 1. In the first case $\mathbb{P}(A) = \mathbb{E}f = 0$ and in the second it is one.

Conversely, suppose that T is ergodic and let f be a T -invariant function. Setting $F(x) = \mathbb{P}(f \leq x)$ to be the distribution function of f , we must have $F(x) \in \{0, 1\}$ for all x , since $\{f \leq x\}$ is T -invariant. However $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$, implying that $F(x) = \mathbf{1}_{[c, \infty)}$ for some $c \in \mathbb{R}$. It follows that $f = c$ with probability one. $\square$

Examples.

1. Let $\Omega = \{z \in \mathbb{C} : |z| = 1\}$ and for $\theta \in \mathbb{R}$, $T(z) = e^{i\theta}z$. Then T preserves Lebesgue measure on Ω (with the Borel sigma-algebra). Actually if $\theta = 2\pi\alpha$ then T is ergodic if and only if α is irrational. See Billingsley, p. 316.

LECTURE 42

2. Let $\mathbb{P}$ be an i.i.d. product measure on $\mathbb{R}^{\mathbb{N}}$. Then $\mathbb{P}$ is ergodic for the shift $T(x_1, x_2, \dots) = (x_2, x_3, \dots)$.

Proof. The proof is similar to that of the Hewitt-Savage zero-one law. Given A , a shift-invariant event, we can find (by the π - λ theorem – you proved this in homework) a sequence of cylinder events A_n depending on the first n coordinates such that

$$\mathbb{P}(A \Delta A_n) \rightarrow 0 . \tag{45}$$

Approximating by the A_n 's helps because shift-invariance of A implies approximate shift-invariance of A_n , and a cylinder event can only be shift-invariant if it is trivial, giving that A is trivial.

For the proof, if we write T for the shift, then A_n and $T^n A_n$ are independent, so

$$\begin{aligned} |\mathbb{P}(A)^2 - \mathbb{P}(A)| &= |\mathbb{P}(A)\mathbb{P}(T^n A) - \mathbb{P}(A \cap T^n A)| \\ &\leq |\mathbb{P}(A)\mathbb{P}(T^n A) - \mathbb{P}(A_n)\mathbb{P}(T^n A_n)| \\ &\quad + |\mathbb{P}(A_n \cap T^n A_n) - \mathbb{P}(A \cap T^n A)| \end{aligned}$$

By shift-invariance of $\mathbb{P}$, we also have

$$\mathbb{P}(T^n A_n \Delta T^n A) \rightarrow 0 .$$

This and (45) implies that the two terms above converge to 0. That is $\mathbb{P}(A)^2 = \mathbb{P}(A)$ and so $\mathbb{P}(A) = 0$ or 1 . $\square$

19.1 Ergodic theorem

The definition of ergodic for sequences is a direct translation of the one above.

Definition 19.5. *A sequence (X_n) of random variables is ergodic if it is stationary and every shift-invariant $A \in \sigma(X_n : n \geq 1)$ has probability zero or one.*

In this definition, shift-invariant means the following. Any $A \in \sigma(X_n : n \geq 1)$ can be written as

$$A = \{\omega : (X_1(\omega), X_2(\omega), \dots) \in H\}$$

for some Borel $H \subset \mathbb{R}^{\mathbb{N}}$. Then A is shift-invariant if H is: if $T : \mathbb{R}^{\mathbb{N}} \rightarrow \mathbb{R}^{\mathbb{N}}$ is given by

$$T(x_1, x_2, \dots) = (x_2, x_3, \dots)$$

then $H = T^{-1}(H)$.

The ergodic theorem is a law of large numbers in this more general setting. The proof comes directly from a lecture of Y. Peres on a simple proof of the ergodic theorem. It is called “Intuitive proofs of the ergodic theorem” and can be found on the microsoft website.

Theorem 19.6 (Birkhoff’s ergodic theorem). *Define $S_l(k) = \sum_{j=0}^{l-1} X_{k+j}$ and $A_l(k) = \frac{1}{l} S_l(k)$. If $\mathbb{E}|X_1| < \infty$ and (X_n) is stationary then*

$$A_n(1) \rightarrow \mathbb{E}[X_1 \mid \mathcal{I}] \text{ almost surely .}$$

There are really two parts to the statement. The first is that the averages converge and the second is that the limit is $\mathbb{E}[X_1 \mid \mathcal{I}]$. Here $\mathcal{I}$ is the invariant sigma-algebra. In the sequence setting, this is the set of events that are invariant under shift. That is, if $\Phi : \Omega \rightarrow \mathbb{R}^{\mathbb{N}}$ is the map

$$\Phi(\omega) = (X_1(\omega), X_2(\omega), \dots)$$

then if we define $T : \mathbb{R}^{\mathbb{N}} \rightarrow \mathbb{R}^{\mathbb{N}}$ as the left-shift and $\mathcal{I}'$ the T -invariant sigma-algebra of $\mathbb{R}^{\mathbb{N}}$, then $\mathcal{I}$ is $\{\Phi^{-1}A : A \in \mathcal{I}'\}$.

In the ergodic case, $\mathcal{I}$ is trivial and so

$$A_n(1) \rightarrow \mathbb{E}X_1 \text{ almost surely .}$$

Proof. This proof is based on Kamae, Katznelson-Weiss and Keane and we will show the ergodic case. It is enough to consider $X_j \geq 0$ for all j because in general we can separate into positive and negative parts:

$$X_j = (X_j)_+ - (X_j)_- \text{ and } A_n(1) = \frac{1}{n} \left(\sum_{j=0}^{n-1} (X_{1+j})_+ - \sum_{j=0}^{n-1} (X_{1+j})_- \right) .$$

So if we prove the result for nonnegative random variables, we can apply it to $(X_j)_+$ and $(X_j)_-$ and retrieve the result for X_j .

Let

$$\bar{A} = \limsup_n A_n(1) \in [0, \infty] .$$

Note that $\bar{A}$ is a shift-invariant random variable: $S_n(1) = S_{n-1}(2) + X_1$ and so

$$\limsup_n A_n(1) = \limsup_n \frac{1}{n} S_n(1) = \limsup_n \frac{1}{n} [S_{n-1}(2) + X_1] = \limsup_n A_n(2) ,$$

By ergodicity, then, $\bar{A}$ is constant almost surely. We will later prove that $\bar{A} < \infty$, but for now we must accept the possibility that $\bar{A} = \infty$. Fix any

$$\alpha < \bar{A}$$

and define

$$l(k) = \min\{l : A_l(k) \geq \alpha\} .$$

(This is the first time that the averages exceed α .)

We first explore a very special case. The general case will be an extension of the following argument.

Case 1. Suppose that $l(k) \leq L$ for all k , where L is a constant. We will just take the partial sum up to index n and cover the interval $[1, n]$ by consecutive subintervals on which the averages exceed α . We will begin covering starting at index 1 and then stop the first time we exceed index n . By assumption, all the subintervals will have length at most L . The endpoints of our subintervals will be k_i and the lengths will be $l(k_i)$:

$$k_1 = 1 \text{ and } k_{i+1} = k_i + l(k_i) \text{ with } m = \max\{i \geq 1 : k_i \leq n\} .$$

Then by non negativity and definition of $l(k)$, we obtain the ‘‘heart of the proof’’

$$\sum_{j=1}^n X_j \geq \sum_{i=1}^m S_{l(k_i)}(k_i) \geq \sum_{i=1}^m l(k_i) \alpha \geq (n - L) \alpha .$$

Here we have just summed the X_j 's on the subintervals and on each one, the sum must exceed $l(k_i)\alpha$. We stop at the last one, $l(k_m)$, and this last endpoint must come within distance L of n , by assumption.

Now $\liminf_n \frac{1}{n} \sum_{j=1}^n X_j \geq \alpha$. This holds for all $\alpha < \bar{A}$, so $\liminf_n A_n(1) \geq \limsup_n A_n(1)$, proving existence of the limit. We will not finish this case, as it was more for motivation.

Next time we will do the general case. $\square$

LECTURE 43

Now we continue with the proof of the general case of the ergodic theorem.

Proof. Case 2. In the general case, we want to almost bound $l(k)$, so given $\epsilon > 0$ pick L such that

$$\mathbb{P}(l(k) > L) < \epsilon .$$

To reduce back to the previous case, we will now modify the variables. We look at the sequence $(X_1, X_2, \dots)$ and for each k , if $l(k) \leq L$ (that is, starting from this variable, we do not need to wait too long for the average to exceed α), we leave X_k as it is, and if $l(k) > L$, we set it equal to α :

$$X_k^* = \begin{cases} X_k & \text{if } l(k) \leq L \\ \alpha & \text{if } l(k) > L \end{cases} .$$

We can write $X_k^* = X_k + Z_k$, where

$$Z_k = \begin{cases} 0 & \text{if } l(k) \leq L \\ \alpha - X_k & \text{if } l(k) > L \end{cases} .$$

Notice here that $Z_k \geq 0$. To see why, suppose that $l(k) > L$. Then $l(k) > 1$ and so by definition, $X_k = A_1(k) < \alpha$, giving $Z_k = \alpha - X_k > 0$.

Writing $A_l^*(k) = \frac{1}{l} \sum_{j=0}^{l-1} X_{j+k}^*$ and $l^*(k) = \min\{l : A_l^*(k) \geq \alpha\}$, then

$$l^*(k) = \begin{cases} l(k) & \text{if } l(k) \leq L \\ 1 & \text{if } l(k) > L \end{cases} . \quad (46)$$

Indeed, if $l(k) > L$ then $X_k^* = \alpha$, and so $l(k)^* = 1 \leq L$. Otherwise $l(k) \leq L$ and so $A_{l(k)}(k) \geq \alpha$ with $A_j(k) < \alpha$ for $j < l(k)$. Now $A_{l(k)}^*(k) \geq A_{l(k)}(k) \geq \alpha$, giving $l^*(k) \leq l(k)$. On the other hand, if $l^*(k) < l(k)$ then $A_{l^*(k)}(k) \geq \alpha$ while $A_{l(k)^*}(k) < \alpha$. This means that there is some $j \leq l^*(k)$ such that $X_{k+j} \neq X_{k+j}^*$ and so $l(k+j) > L$. In particular, $l(k+j) > l(k) - j$ and so

$$A_{l(k)}(k) = \frac{1}{k} (S_j(k) + S_{l(k)-j}(k+j)) \leq \frac{1}{2} (A_j(k) + A_{l(k)-j}(k+j)) < \alpha ,$$

contradicting the definition of $l(k)$. This shows $l^*(k) \geq l(k)$ and consequently (46).

Now X_j^* satisfies the assumptions of the last case, so by exactly the same argument as before, we get the second main inequality

$$\sum_{j=1}^n X_j^* \geq (n - L)\alpha \text{ for all } n. \quad (47)$$

To derive this inequality, we do not even need to know that the X_j^* 's are ergodic or integrable (see the previous proof). We only need that they are nonnegative. Taking expectation and using stationarity,

$$(n - L)\alpha \leq n\mathbb{E}X_1^* = n(\mathbb{E}X_1 + \mathbb{Z}_1).$$

Because $Z_1 \leq \alpha$ and is only nonzero when $l(1) > L$ (and this event has probability at most ϵ), we obtain

$$(n - L)\alpha \leq n(\mathbb{E}X_1 + \alpha\epsilon),$$

or $\mathbb{E}X_1 \geq (1 - \epsilon)\alpha$. This is true for all $\alpha < \bar{A}$ and $\epsilon > 0$, so

$$\mathbb{E}X_1 \geq \bar{A},$$

proving that $\bar{A} < \infty$.

Now we apply this result also to the sequence (Z_n) . It is integrable, since it takes values 0 or $\alpha - X_k$ (but only when $X_k < \alpha$) and it is still stationary. You must only verify that it is ergodic (this follows from a general fact that you can check – a stationary sequence which is a function of an ergodic sequence is still ergodic). So also

$$\alpha\epsilon \geq \mathbb{E}Z_1 \geq \limsup_n \frac{1}{n} \sum_{j=1}^n Z_j.$$

Returning to (47),

$$\frac{1}{n} \left(\sum_{j=1}^n X_j + \sum_{j=1}^n Z_j \right) \geq \alpha(1 - L/n).$$

So

$$\alpha \leq \liminf_n \frac{1}{n} \sum_{j=1}^n X_j + \limsup_n \frac{1}{n} \sum_{j=1}^n Z_j \leq \liminf_n A_n(1) + \alpha\epsilon.$$

This gives

$$\liminf_n A_n(1) \geq \alpha(1 - \epsilon)$$

and since ϵ is arbitrary and α is any number $< \bar{A}$,

$$\liminf_n A_n(1) \geq \bar{A}, \text{ so } \liminf_n A_n(1) = \limsup_n A_n(1)$$

and the limit exists.

To prove that $\bar{A} = \mathbb{E}X_1$, we have already shown the inequality $\geq$. For $\leq$, note that if the X_j 's were bounded, we could use dominated convergence for equality. In the general case, we apply the result we already have to $\min\{X_j, M\}$ for $M \geq 0$:

$$\bar{A} \geq \lim_n \frac{1}{n} \sum_{j=1}^n \min\{X_j, M\} = \mathbb{E} \min\{X_1, M\} .$$

Take $M \rightarrow \infty$ for $\bar{A} \geq \mathbb{E}X_1$. □

19.2 Structure of ergodic measures

Of the many characterizations of ergodic measures, one of the more useful ones exhibits them as extremal invariant measures. Define

For a given (Ω, Σ) and $T : \Omega \rightarrow \Omega$ measurable, we define

$$\mathcal{M}_T = \{T\text{-invariant probability measures on } (\Omega, \Sigma)\} .$$

Recall that a subset S of a vector space V over $\mathbb{R}$ is convex if $\lambda v + (1 - \lambda)w \in S$ whenever $v, w \in S$ and $\lambda \in (0, 1)$. A point $v \in S$ is called an *extreme point* of S if it cannot be written as a convex combination of other elements of S ; that is,

$$v = \lambda w_1 + (1 - \lambda)w_2 \text{ for } \lambda \in (0, 1) \text{ and } w_1, w_2 \in S \Rightarrow w_1 = w_2 .$$

Proposition 19.7. *The set $\mathcal{M}_T$ is convex and its extreme points are the ergodic probability measures. Furthermore distinct ergodic measures are mutually singular. That is, if μ, ν are ergodic and not equal, there exists $A \in \Sigma$ such that $\mu(A) = 1 = \nu(A^c)$.*

Proof. If $\mu_1, \mu_2 \in \mathcal{M}_T$ and $\lambda \in [0, 1]$, then for $A \in \Sigma$, setting $\mu = \lambda\mu_1 + (1 - \lambda)\mu_2$,

$$\mu(T^{-1}A) = \lambda\mu_1(T^{-1}A) + (1 - \lambda)\mu_2(T^{-1}A) = \lambda\mu_1(A) + (1 - \lambda)\mu_2(A) = \mu(A) .$$

Next assume that μ is not ergodic. Then there exists $A \in \mathcal{I}$ such that $\mu(A) \in (0, 1)$. Then we can write

$$\mu(\cdot) = \mu(\cdot \mid A)\mu(A) + \mu(\cdot \mid A^c)\mu(A^c) .$$

Both $\mu(\cdot \mid A)$ and $\mu(\cdot \mid A^c)$ are T -invariant, and μ is a non-trivial convex combination of them, so it is not extreme.

Conversely suppose that μ is ergodic. If we write $\mu = \lambda\mu_1 + (1 - \lambda)\mu_2$ for $\lambda \in (0, 1)$ and $\mu_1, \mu_2 \in \mathcal{M}_T$, then for any $A \in \mathcal{I}$,

$$\lambda\mu_1(A) + (1 - \lambda)\mu_2(A) = \mu(A) \in \{0, 1\} .$$

Since $\lambda \in (0, 1)$, this implies that $\mu_1(A) = \mu_2(A)$ for all $A \in \mathcal{I}$. To show they agree on Σ , let $A \in \Sigma$ and apply the ergodic theorem to μ : for some set $E \in \Sigma$ with $\mu(E) = 1$,

$$\frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_A(T^k \omega) \rightarrow \mu(A) \text{ for } \omega \in E .$$

It follows that also $\mu_1(E) = \mu_2(E) = 1$ and so this convergence occurs μ_1 -almost surely and μ_2 -almost surely. But then by the dominated convergence theorem, for $i = 1, 2$,

$$\mu_i(A) = \lim_n \mathbb{E}_{\mu_i} \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_A(T^k \omega) = \mu(A) ,$$

meaning $\mu_1 = \mu_2$.

Last we have shown above that if two measures $\mu, \nu \in \mathcal{M}_T$ agree on $\mathcal{I}$ then they are equal. So if $\mu \neq \nu$ we can find $A \in \mathcal{I}$ such that $\mu(A) \neq \nu(A)$. If these measures are ergodic, then one of these probabilities must be 0 and one must be 1, giving mutual singularity. $\square$